Earth

Evolution of a Habitable World

Second edition

Fully updated throughout, including revised illustrations and new images from NASA missions, this new edition provides an overview of Earth's history from a planetary science perspective, for undergraduates in earth science, planetary science, and astronomy. The evolution of the Earth is described in the context of what we know about other planets and the cosmos at large, from the origin of the cosmos to the processes that shape planetary environments, and from the origins of life to the inner workings of cells.

Key features

- Integrates astronomy, earth science, planetary science, and astrobiology to give students the whole picture of how the Earth has come to its present state.
- Presents concepts in nontechnical language and avoids mathematical treatments where possible, allowing students to grasp concepts without wading through complex maths.
- New end-of-chapter summaries and questions allow students to check their understanding and critical thinking is emphasized to encourage students to explore ideas scientifically for themselves.

Jonathan I. Lunine is the David C. Duncan Professor in the Physical Sciences at Cornell University. His research interests center broadly on planetary origin and evolution, in our solar system and around other stars. He works as an interdisciplinary scientist on the Cassini mission to Saturn, and on the James Webb Space Telescope, and is also a co-investigator on the Juno mission, which launched for Jupiter in August 2011. Dr. Lunine is the author of over 230 scientific papers and besides the first edition of this book (Cambridge University Press, 1999), he has also written *Astrobiology: A Multidisciplinary Approach* (Pearson Addison-Wesley, 2005). He is a member of the US National Academy of Sciences, and a fellow of the American Association for the Advancement of Science and the American Geophysical Union.

Praise for this book:

“review quote to come review quote to come review quote to come review quote to come review quote to come review quote to come review quote to come review quote to come review quote to come”
Reviewer 1, somewhere

“review quote to come review quote to come review quote to come review quote to come review quote to come review quote to come review quote to come review quote to come review quote to come”
Reviewer 2, somewhere

“review quote to come review quote to come review quote to come review quote to come review quote to come review quote to come review quote to come review quote to come review quote to come”
Reviewer 3, somewhere

“review quote to come review quote to come review quote to come review quote to come review quote to come review quote to come review quote to come review quote to come review quote to come”
Reviewer 4, somewhere

Earth Evolution of a Habitable World

Second edition

Jonathan I. Lunine
Cornell University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521850018

First edition © Cambridge University Press 1999
Second edition © Jonathan I. Lunine 2013

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without the written
permission of Cambridge University Press.

First published 1999
Reprinted 2000
Second edition 2013

Printed in the United Kingdom at the University Press, Cambridge

A catalogue record for this publication is available from the British Library

Library of Congress Cataloguing in Publication data

ISBN 978-0-521-85001-8 Hardback
ISBN 978-0-521-61519-8 Paperback

Additional resources for this publication at www.cambridge.org/lunine

Cambridge University Press has no responsibility for the persistence or
accuracy of URLs for external or third-party internet websites referred to
in this publication, and does not guarantee that any content on such
websites is, or will remain, accurate or appropriate.

CONTENTS

Preface *page* ix

Part I The astronomical planet: Earth's place in the cosmos

1 An introductory tour of Earth's cosmic neighborhood 3
1.1 Ancient attempts to determine the scale of the cosmos 3
1.2 Brief introduction to the solar system 4
Summary 6
Questions 7
Further reading 7
Reference 7

2 Largest and smallest scales 9
Introduction 9
2.1 Scientific notation 9
2.2 Motions of Earth in the cosmos 9
2.3 Cosmic distances 13
2.4 Microscopic constitution of matter 17
Summary 22
Questions 22
General reading 23
References 23

3 Forces and energy 25
Introduction 25
3.1 Forces of nature 25
3.2 Radioactivity 29
3.3 Conservation of energy, and thermodynamics 29
3.4 Electromagnetic spectrum 30
3.5 Abundances in the Sun 31
Summary 33
Questions 34
References 34

4 Fusion, fission, sunlight, and element formation 35
Introduction 35
4.1 Stars and nuclear fusion 35
4.2 Element production in the Big Bang 38
4.3 Element production during nuclear fusion in stars 38
4.4 Production of other elements in stars: *s*, *r*, and *p* processes 39
4.5 Nonstellar element production 41
4.6 Element production and life 41
Summary 42
Questions 42
References 43

Part II The measurable planet: tools to discern the history of Earth and the planets

5 Determination of cosmic and terrestrial ages 47
Introduction 47
5.1 Overview of age dating 47
5.2 The concept of half-life 47
5.3 Carbon-14 dating 49
5.4 Measurement of parents and daughters: rubidium–strontium 50
5.5 Fission track dating 52
5.6 Caveat emptor 52
Summary 53
Questions 53
General reading 53
References 54

6 Other uses of isotopes for Earth history 55
Introduction 55
6.1 Stable isotopes, seafloor sediments, and climate 55
6.2 A possible temperature history of Earth from cherts 57
Summary 59

Questions 60
General reading 60
References 60

7 Relative age dating of cosmic and terrestrial events: the cratering record 61
Introduction 61
7.1 Process of impact cratering 61
7.2 Using craters to date planetary surfaces 62
7.3 Cratering on planetary bodies with atmospheres 68
7.4 Impactors through time 70
Summary 70
Questions 70
References 71

8 Relative age dating of terrestrial events: geologic layering and geologic time 73
Introduction 73
8.1 Catastrophism versus uniformitarianism 73
8.2 Estimating the age of Earth, without radioisotopes 73
8.3 Geologic processes and their cyclical nature 74
8.4 Principles of geologic succession 76
8.5 Fossils 77
8.6 Radioisotopic dating of Earth rocks 79
8.7 Geologic timescale 79
8.8 A grand sequence 80
8.9 The geologic timescale as a map 81
Summary 81
Questions 81
General reading 82
References 82

9 Plate tectonics: an introduction to the process 83
Introduction 83
9.1 Early evidence for and historical development of plate tectonics 83
9.2 Genesis of plate tectonics after World War II 84
9.3 The basic model of plate tectonics 87
9.4 Past motions of the plates and supercontinents 91
9.5 Driving forces of plate motions 94
9.6 An end to techniques and the start of history 95
Summary 95
Questions 95
General reading 95
References 96

Part III The historical planet: Earth and solar system through time

10 Formation of the solar system 99
Introduction 99
10.1 Timescale of cosmological events leading up to solar system formation 99
10.2 Formation of stars and planets 100
10.3 Primitive material present in the solar system today 105
10.4 The search for other planetary systems 107
10.5 Summary of planet formation 110
Summary 111
Questions 111
General reading 111
References 112

11 The Hadean Earth 113
Introduction 113
11.1 Bulk composition of the planets 113
11.2 Internal structure of Earth 117
11.3 Accretion: the building up of planets 120
11.4 Early differentiation after accretion 121
11.5 Radioactive heating 122
11.6 Formation of an iron core 123
11.7 Formation of the Moon 123
11.8 Origin of Earth's atmosphere, ocean, and organic reservoir 125
11.9 The Late Heavy Bombardment 126
11.10 From the Hadean into the Archean: formation of the first stable continental rocks 127
Summary 128
Questions 128
General reading 129
References 129

12 The Archean eon and the origin of life I Properties of and sites for life 131
Introduction 131
12.1 Definition of life and essential workings 131
12.2 The basic unit of living organisms: the cell 135
12.3 Energetic processes that sustain life 136
12.4 Other means of utilizing energy 136
12.5 Elemental necessities of life: a brief examination 138
12.6 Solar system sites for life 140
Summary 146
Questions 147
General reading 148
References 148

13 The Archean eon and the origin of life II Mechanisms 149
Introduction 149
13.1 Thermodynamics and life 149
13.2 The raw materials of life: synthesis and the importance of handedness 151
13.3 Two approaches to life's origin 152
13.4 The vesicle approach and autocatalysis 152
13.5 The RNA world: a second option 154
13.6 The essentials of a cell and the unification of the two approaches 156
13.7 The Archean situation 158
Summary 159
Questions 159
General reading 160
References 160

14 The first greenhouse crisis: the faint young Sun 161
Introduction 161
14.1 The case for an equable climate in the Archean 161
14.2 The faint young Sun 161
14.3 The greenhouse effect 162
14.4 Primary greenhouse gases 164
14.5 Implications for Earth during the faint young Sun era 164
14.6 Paleosols and the carbon dioxide abundance 166
14.7 Carbon dioxide cycling and early crustal tectonics 167
14.8 A balance unique to Earth, and a lingering conundrum 170
Summary 171
Questions 171
General reading 172
References 172

15 Climate histories of Mars and Venus, and the habitability of planets 173
Introduction 173
15.1 Venus 173
15.2 Mars 178
15.3 Was Mars really warm in the past? 181
15.4 Putting a Martian history together 184
15.5 Implications of Venusian and Martian history for life elsewhere 184
15.6 The finite life of our biosphere 185
Summary 186
Questions 186
General reading 187
References 187

16 Earth in transition: from the Archean to the Proterozoic 189
Introduction 189
16.1 Abundances of the elements in terrestrial rocks 189
16.2 Mineral structure 190
16.3 Partial melting and the formation of basalts 191
16.4 Formation of andesites and granites 192
16.5 Formation of protocontinents in the Archean 195
16.6 The Archean–Proterozoic transition 196
16.7 After the Proterozoic: modern plate tectonics 197
16.8 Venus: an Earth-sized planet without plate tectonics 198
16.9 Water and plate tectonics 199
16.10 Continents, the Moon, and the length of Earth's day 200
16.11 Entree to the modern world 201
Summary 201
Questions 201
General reading 202
References 202

17 The oxygen revolution 203
Introduction 203
17.1 The modern oxygen cycle 203
17.2 The balance of oxygen with and without life 205
17.3 Limits on oxygen levels on early Earth 205
17.4 History of the rise of oxygen 207
17.5 Balance between oxygen loss and gain 207
17.6 Reservoirs of oxygen and reduced gases 208
17.7 History of oxygen on Earth 209
17.8 Shield against ultraviolet radiation 210
17.9 Onset of eukaryotic life 211
Summary 212
Questions 213
References 213

18 The Phanerozoic: flowering and extinction of complex life 215
Introduction 215
18.1 Evolution 217
18.2 Ediacaran–Cambrian revolution 220
18.3 Mass extinction events in the Phanerozoic 223
18.4 Cretaceous–Tertiary extinction 223
18.5 A global view of Earth's history so far 227
Summary 228
Questions 228
General reading 229
References 229

19 Climate change across the Phanerozoic 231
Introduction 231
19.1 The supercontinent cycle 231
19.2 Effects of continental break ups and collisions 233
19.3 Evidence of ice ages on Earth 233

19.4 Causes of the ice ages 234
19.5 Cretaceous climate 235
19.6 The great Tertiary cool down 237
19.7 Causes of the Pleistocene ice age and its oscillations 239
19.8 Saved from instability: Earth's versus Mars' orbital cycle 241
19.9 Effects of the Pleistocene ice age: a preview 242
Summary 242
Questions 243
References 243

20 Toward the age of humankind 245
Introduction 245
20.1 Pleistocene setting 245
20.2 The vagaries of understanding human origins 245
20.3 Humanity's taxonomy 246
20.4 The first steps: Australopithecines 246
20.5 The genus *Homo*: Out of Africa I 247
20.6 Out of Africa II 248
20.7 Final act: Neanderthals and an encounter with our humanity 249
20.8 This modern world 253
Summary 254
Questions 254
General reading 254
References 254

Part IV The once and future planet

21 Climate change over the past few hundred thousand years 259
Introduction 259
21.1 The record in ice cores 259
21.2 Climate from plant pollen and packrat midden studies 261
21.3 Tree rings 264
21.4 Climate variability in the late Holocene 266
21.5 The Younger Dryas: a signpost for the oceanic role in climate 267
21.6 Into the present 268
Summary 268
Questions 269
General reading 269
References 269

22 Human-induced global warming 271
Introduction 271
22.1 The records of CO_2 abundance and global temperatures in modern times 271
22.2 Modeling the response of Earth to increasing amounts of greenhouse gases 273
22.3 Predicted effects of global warming 276
22.4 The difficulty of proof: weather versus climate 280
22.5 Role of the oceans in Earth's climate 281
22.6 Global warming: a long-term view 283
22.7 Postscript: human effects on the upper atmosphere – ozone depletion 284
Summary 284
Questions 285
General reading 285
References 285

23 Limited resources: the human dilemma 287
Introduction 287
23.1 The expanding human population 287
23.2 Prospects for agriculture 288
23.3 Energy resources 289
23.4 Economically important minerals 292
23.5 Pollution 294
23.6 Can we go back? 295
Summary 296
Questions 296
General reading 296
References 297

24 Coda: the once and future earth 299

Index 301

Color plate section is between pages xx and xx.

PREFACE

When the first edition of this book was published some 15 years ago, astrobiology was not recognized as a separate academic discipline, and few universities and colleges offered courses in the subject per se. But the question of what makes a planet capable of sustaining life, and whether inhabited planets exist in large numbers in the cosmos, was long a popular draw for courses in planetary science, geology, and astronomy. I wrote *Earth: Evolution of a Habitable World* so as to encourage instructors of freshmen and sophomore non-science majors to take a consciously planetary bent in covering how our home planet came to be, its place in the overall evolution of the cosmos, how it became habitable *and* inhabited, and how life and the environment evolved together (sometimes coupled, sometimes not) to the present day. And in closing with chapters on human-induced global warming and depletion of resources, I wished to provide a "cosmic perspective" via the rest of the book to some very down-to-Earth problems. In the breadth of topics and perspective I took in writing it, *Earth* was alone in its chosen subject area, with only a few notable exceptions.

Today astrobiology is a thriving academic field with a daunting number and variety of textbooks on the subject. In preparing a revised edition I considered making the book more consciously astrobiological, either by aligning the contents more closely with the typical survey treatment – or by simply adding the word "Astrobiology" to the title. But neither option seemed to me to do justice to the main theme of the text, which remains the story of our planet Earth from its cosmic beginnings to the present-day practical dilemmas our success as a technological species has brought us. The astute instructor or student will be able to figure out that the book is suitable for a course in an astrobiology program, just as one might understand that a textbook entitled *Classical Mechanics* is suitable for covering part of a physics curriculum. The level remains the same, parts have been updated or rewritten, new figures included, and quiz questions expanded. As before, the book also will be useful to those who are not enrolled in courses but want to learn something of Earth's history from a planetary perspective. However, I am well aware that there is much more competition today for both the student and interested layman, and I can only hope that this particular narrative finds its niche within the plethora of astrobiology books.

The first edition of the book was prepared when I was on the faculty of the Lunar and Planetary Laboratory, Department of Planetary Sciences, The University of Arizona. I remain forever in debt for the help, encouragement, and contributions of my colleagues there. The second edition was prepared while I was on leave of absence to the University of Rome Tor Vergata, Rome, Italy, and completed here at Cornell University where I now teach; both of these institutions provided assistance and encouragement. Likewise I thank Phil Eklund, who as with the first edition provided stimulating comments, suggestions, and figure ideas. My wife Cynthia Lunine illustrated the first edition but other commitments prevented her from preparing new ones for the revised edition. Nonetheless the clarity and attractiveness of style are the direct result of her work, for which I am deeply grateful.

PART I

The astronomical planet: Earth's place in the cosmos

1

An introductory tour of Earth's cosmic neighborhood

1.1 Ancient attempts to determine the scale of the cosmos

The science of astronomy developed in many different cultures and from many different motivations. Because, even in cities of the preindustrial world, the stars could be seen readily at night, the pageant of the sky was an inspiration for, and embodiment of, the myths and legends of almost all cultures. Some people tracked the fixed stars and moving planets with great precision, some for agricultural purposes (the ancient Egyptians needed to prepare for the annual flooding of the Nile River Valley) and more universally to attempt to predict the future. The regularity of the motions of the heavens was powerfully suggestive of the notion that history itself was cyclical, and hence predictable. The idea of human history linked to celestial events remains with us today as the practice of astrology. In spite of a lack of careful experimental tests, or demonstrated physical mechanisms, this powerfully attractive belief system is pursued widely with varying amounts of seriousness, extending in the early 1980s to the level of the presidency of the United States.

Although ancient understanding of the nature of the cosmos varied widely and was usually a reflection of particular mythologies of a given culture, the classical Greeks distinguished themselves by their (often successful) attempts to use experiment and deduction to learn about the universe. Some Greek philosophers understood the spherical nature of Earth and something of the scale of nearby space. Aristotle, in the fourth century BC, correctly interpreted lunar eclipses as being due to the shadow of Earth projected on the surface of the Moon. By noting that the shadow was rounded, he deduced that Earth must be spherical; in fact, another acceptable shape based on that one observation is a disk (Figure 1.1). Others, such as Plato, had much earlier endorsed a spherical shape on aesthetic grounds.

Eratosthenes, who lived in the third century BC, made a remarkably accurate determination of the size of our planet without having to travel too far. He used the observation that at high noon on summer solstice (June 21 in our calendar, when the Sun reaches its northernmost point in the sky of Earth), the Sun was directly overhead at a site in Syene (now Aswan), Egypt, because no shadow could be seen in the vertical well shaft. Eratosthenes lived in Alexandria, due north of Syene, and there he could observe that the Sun cast a shadow at noon on that same date of June 21 (Figure 1.1).

What did this mean? If Earth were a sphere, then different people standing at different locations on Earth at the same time would see the Sun in different parts of the sky. By measuring as an angular distance in the sky, the change in the position of the Sun from one place to another and knowing the distance between the two stations, one could then by a simple calculation work out the circumference of the whole globe. In his home city, Eratosthenes carefully measured the size of a shadow cast by an obelisk of known height, at the same time on the same day that no noontime shadow occurred at Syene. The angular position of the Sun, from the size of the shadow at Alexandria, gave an angle of 7.2 degrees between the position of the Sun at the two stations, or one-fiftieth of the entire angular extent of the sky (which by definition surrounds our globe and therefore extends over 360 degrees). Therefore, Earth's circumference, he knew, must be 50 times the distance between Syene and Alexandria.

The distance was, however, known only approximately from the number of days it took a camel to travel between the two towns and the distance a typical camel walks in a day. Furthermore, to compare the result with the value we know today, the units of measurement used by the Greeks must be converted to modern ones, which is also an uncertain exercise. In modern units, the Syene–Alexandria distance is 570 miles, or 918 kilometers (km), and hence Eratosthenes' experiment yields an Earth circumference of 46,000 km, just 12% too large. This represents an extraordinary achievement, 2,300 years before human beings could view the round globe of Earth from space.

Not everything about the cosmos that the Greek philosophers deduced or inferred came out right. The most celebrated mistake was that of Ptolemy, who lived 400 years after Eratosthenes and is associated most closely with the cosmological system in which the Sun and the planets (in fact, the whole cosmos) were thought to orbit Earth. However, this was just the penultimate round in a long debate on the topic: Aristarchus of Samos, a generation before Eratosthenes, put the Sun at the center with

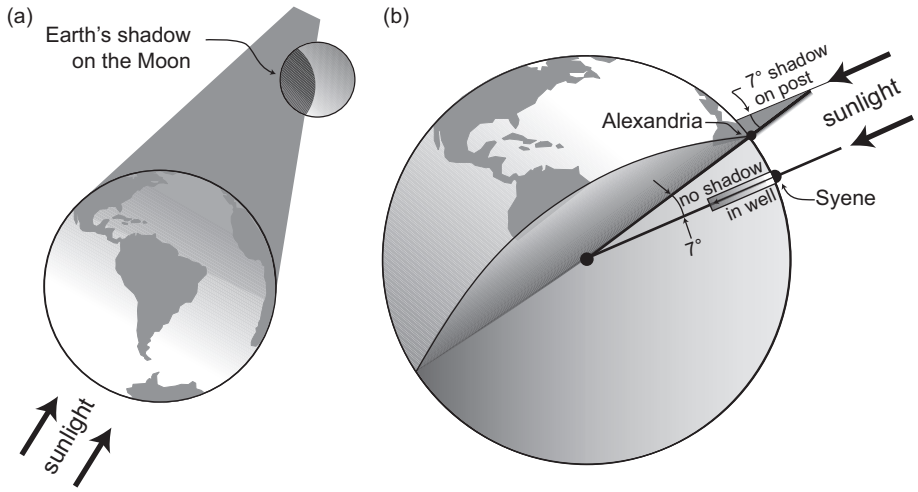


Figure 1.1 Two ancient Greek observations of the cosmos: (a) Aristotle's determination of Earth's sphericity via a lunar eclipse; (b) Eratosthenes' measurement of the size of Earth. Adapted from Snow (1991).

Earth and the other planets orbiting it. This correct model of the solar system was discredited at the time because the Greeks could not see the stars shift in position as Earth moved from one point in its orbit to the opposite side. In fact, the stars do appear to shift position, in the phenomenon called parallax that we describe later, but they are so far away that the shift cannot be detected with the unaided eye. This the Greeks did not know, and the failed experiment led them down the wrong path of an Earth-centered cosmos that would not finally be discarded until the times of Copernicus and Galileo, over 1,500 years later.

We should not fault the classical Greeks for their wrong interpretations, but should admire their startling successes, which were based on observations unaided by the technologies available at present, coupled with the disciplined logic of inductive and deductive reasoning, which was the foundation of the scientific method. Few of us today could repeat the insights of the handful of extraordinary philosophers who anticipated by many centuries some of the outcomes of the Copernican Revolution. In point of fact, we in the industrialized world still have a mindset in essence of an Earth-centered universe: we think little of the sky, increasingly obscured by the lights of cities and hence unfamiliar to us, unless it is to wonder when the Sun will set today, or what the local newspaper horoscope claims our immediate future will hold.

1.2 Brief introduction to the solar system

The solar system consists of eight major planets, several classes of minor planets, some 166 (as of the beginning of 2006) named natural satellites (or *moons*), and innumerable small bodies, all orbiting the Sun. In 2006 Pluto was "demoted" by the International Astronomical Union from the status of planet to a member of a class of "dwarf planets" that include other members in the region beyond the orbit of Neptune, and the largest bodies in the "main asteroid belt" between Mars and Jupiter. Robotic spacecraft have traversed the distance to the farthest planet in the solar system, some 6 billion km. The distance to the nearest star, Proxima Centauri, is 6,000 times greater; hence, we have no hope of seeing spacecraft reach such targets in the foreseeable future. In view of this, the solar system is our cosmic neighborhood, accessible for study by spacecraft and constituting the setting within which Earth has evolved through time.

Here the solar system is summarized in tutorial form to provide a context for what follows. The information presented is the result of at least three millennia of observations and insights, capped by three decades of intense scientific study from the ground and space. Some of this effort is described in the book, but to present a complete history of the exploration of the solar system would require a separate volume.

Figure 1.2 is a map of the solar system. The eight planets fall roughly into three classes according to their size and composition. The four *terrestrial* planets Mercury, Venus, Earth, and Mars range in diameter from 4,800 km (Mercury) to 12,700 km (Earth). They occupy a small, inner region of the solar system, and are composed of a mixture of rocky and metallic materials.

The four *giant* or *Jovian* planets Jupiter, Saturn, Uranus, and Neptune are substantially bigger than Earth, ranging in diameter from 49,000 km (Neptune) to 142,000 km (Jupiter). They are much farther from the Sun than are the inner planets: Jupiter's distance from the Sun is five times that of Earth's and hence is abbreviated as 5 *astronomical units* (AU); Neptune is 30 AU from the Sun. In terms of common units of distance, Earth lies 150 million km from the Sun, and thus Neptune is more than 4 billion km from the solar system's center.

The giant planets are composed of a mixture of rocky and icy material and varying amounts of gases; Jupiter and Saturn are mostly hydrogen and helium gas whereas Uranus and Neptune are predominantly icy and rocky material with lesser amounts of hydrogen and helium gas. (Rocky and icy material is used here to mean atoms of silicon, magnesium, iron, oxygen, carbon, nitrogen, sulfur, and others that tend to form rocky and icy materials under conditions of normal pressure. Because of the

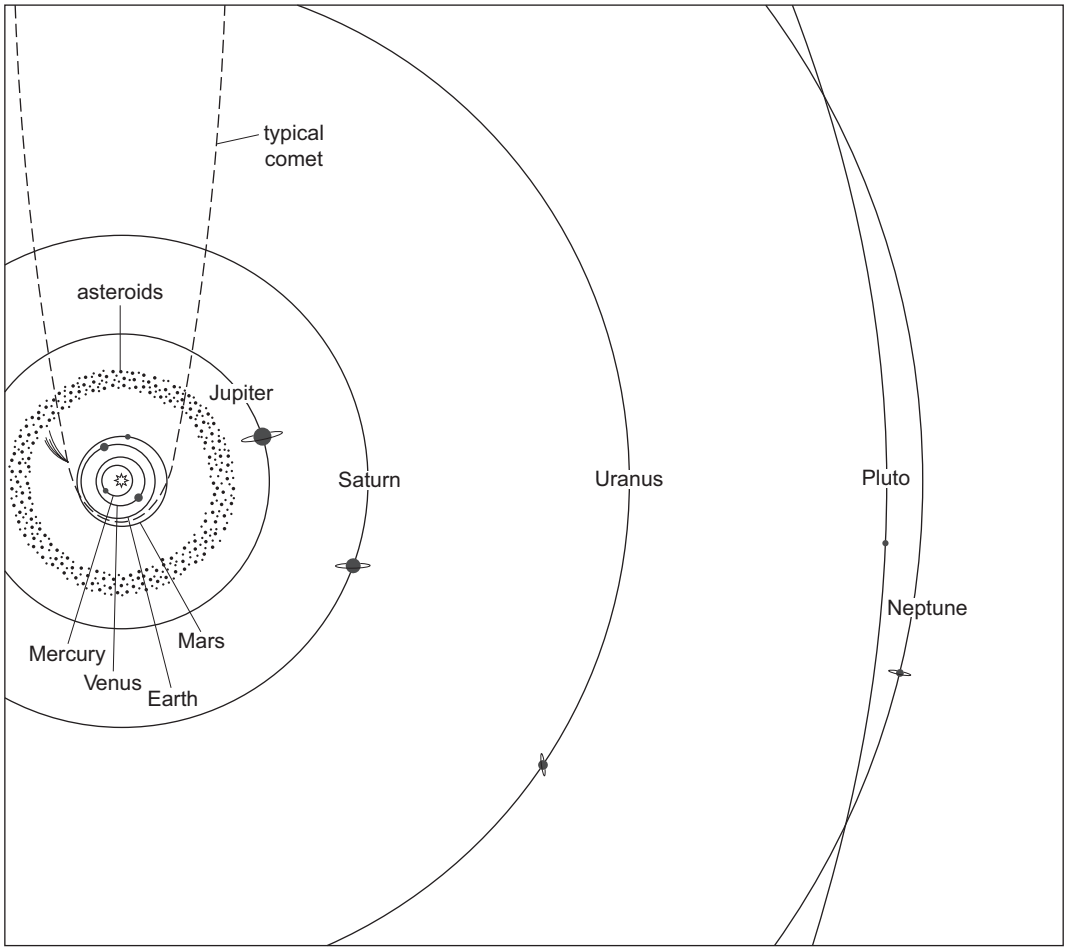


Figure 1.2 Schematic map of our solar system, showing the correct relative sizes of orbits but not of the bodies themselves. Note the small scale of the orbits of terrestrial planets compared to the vast realm of the outer planets. Not shown are the Kuiper Belt beginning just beyond Neptune's orbit and the Oort cloud of comets much further out.

intense pressures deep within these giant planets, much of the icy and rocky material is in atomic form, rather than the molecular form with which we are familiar.)

Six of the planets have moons, as does Pluto and some asteroids. Some moons are small, irregular fragments kilometers across; others – two moons of Jupiter, one of Saturn – are larger than the planet Mercury. The giant planets have multiple satellite systems, some in very regular, circular orbits, which can be considered as miniature solar systems. A class of giant moons, with sizes from that of the Earth's moon upward, include the four Galilean satellites of Jupiter and Saturn's moon Titan. Titan possesses an atmosphere thicker than ours on Earth and sports river channels and perhaps lakes filled or once filled with methane; several other moons have tenuous atmospheres, including our own Moon, which exhibits an extremely rarefied atmosphere of sodium and potassium. All of the planets have atmospheres, though that of Mercury is like our Moon's in being very tenuous.

The four giant planets have ring systems composed of debris from house-sized to dust, which orbits in the equatorial plane of the planet. Saturn's famous ring system is considerably more massive than those of the other major planets. None of the terrestrial planets possesses an organized ring system.

Beyond Neptune lies a part of the solar system poorly explored but, paradoxically, the easiest to see from neighboring stars because of the extensive amount of debris there. The two largest bodies in this region are Pluto and Eris, each about 2,500 km in diameter, and smaller than four of the solar system's moons (Earth's Moon, Jupiter's Ganymede, Callisto, and Saturn's Titan). But they are the largest of a class of debris left over from the formation of the solar system. When Pluto was discovered in 1930 by the American Astronomer Clyde Tombaugh, other bodies of such size beyond Neptune were not known, and hence Pluto was classified as a planet. Today we know that Pluto is a part of the "trans-Neptunian region", or "Kuiper Belt", in which hundreds of other bodies have been individually identified and their orbits mapped. The inner edge of this thick belt of material is defined by the giant planets, whose strong gravitational fields have swept the region from 5 to 30 AU clear of debris and cleaned out lanes within the Kuiper Belt itself. Eris, discovered in 2003, is a bit more massive and larger in size than Pluto. In both size and density (amount of mass per volume in the object), Pluto and Eris are remarkably similar to Triton, the largest moon of Neptune, suggesting this latter to have once been a Kuiper Belt object and further hinting at some sort of natural upper limit to the growth of these bodies.

Well beyond the region of Neptune and the Kuiper Belt lies more icy and rocky material in distant orbits ranging out to perhaps 100,000 AU from the Sun. The presence of such material is inferred from the existence of comets, rock-ice bodies perhaps 1 to 10 km in diameter that come into the inner solar system on highly noncircular, that is, elliptical, orbits. Careful plotting of the paths of comets indicates that most of the orbits originate in an ill-defined shell of material termed the *Oort Cloud*. The comets are the small fraction of Oort Cloud objects that fall inward to the Sun after having been perturbed by close-passing stars. The total number of comet-sized Oort Cloud objects may exceed one trillion.

Remote observation of comets as they pass through the inner solar system suggests that they are accumulations of dust, organic material, water ice, and frozen gases. The Oort Cloud material is thought to have been ejected from the 5- to 30-AU region by the giant planets after their formation and, in addition to comet-sized bodies, both larger and smaller objects may reside in this cloud.

Between the orbits of Mars and Jupiter lie belts of rocky objects known as asteroids. The largest asteroids are several hundred kilometers across; in number and total mass they are minuscule compared to the Oort Cloud and the Kuiper Belt. They are thought to be debris that never formed into a planet because of the proximity of Jupiter, whose gravitational field prevented efficient growth of a large body from smaller ones. Another collection of asteroids crosses the orbit of Earth – the so-called *near-Earth asteroids*, some of which may be old comets that have lost their mantles of ice after many passes by the Sun. Finally, lanes and regions of dust released from comets or asteroids lace the solar system; the precise distribution of this material, some of which can be seen faintly after sunset as the *zodiacal light*, remains somewhat uncertain.

The history of collisions between the numerous bits of small debris and the planets is recorded by the ubiquitous existence of craters throughout the solar system. Even Earth shows the scars, Meteor Crater in Arizona being a famous recent example. As we shall see, impacts may have played key roles in the origin and evolution of life on this planet Earth.

The solar system exhibits several regularities in its structure, which are important in understanding its origin, as we discuss later. All planets orbit the Sun in nearly circular orbits, close to the plane of the Sun's equator. The orbits of Pluto and Eris are more typical of the Kuiper Belt, being *inclined* (tilted relative to the Sun's equator) and *eccentric* (significantly noncircular). All orbits are in the same direction; by convention, they are counterclockwise around the Sun when viewed from above the Sun's *northern hemisphere*. With two exceptions, Venus and Uranus, all planetary spins are in the same, counterclockwise, direction. However, the planetary rotational axes are all tilted relative to their orbital planes by varying degrees.

There is a strong correlation between the properties of the planets and their location in the solar system. The four terrestrial planets, which contain proportionately little water and gases, are closest to the Sun and not very massive compared to the giant planets. From Jupiter outward, solid objects (moons and Pluto) contain significant amounts of water ice and more volatile species. (Here, volatile refers to the tendency for a material to transform from a condensed state to a vapor.) The four giant planets seem to be of two classes, with the more gaseous planets, Jupiter and Saturn, closer to the Sun. These regularities provide clues to the origin of the solar system, but most other planetary systems known to exist around other stars do not exhibit such strict regularities as we discuss in Chapter 10.

Summary

Astronomy arose from the practical and the curious: from the need to keep track of time for planting to the questions of where we came from and whether we are alone in the cosmos. The classical Greeks of 2,500 years ago applied geometry and rigorous thinking to the question of the size of the Earth and distances to the Sun and to the stars. Our cosmic backyard is the solar system, which consists of planets, moons, and numerous smaller bodies all in orbit around a rather commonplace star we call (in English) the Sun. Evidence that the planets formed from accumulation of smaller material comes from the record of craters – holes formed in the surfaces of the solid planets and moons by high speed impacts. The planets of our solar system seem to be well ordered, with rocky planets orbiting close to the Sun and gas giants with icy moons further out, a situation that may not be the norm for planetary systems around other stars.

Questions

1. Consider how you have responded to a controversial scientific or technological issue. Did you try to weigh rationally the pros and cons, or did you respond on the basis of your instincts or emotions? In your own experience, which approach – the rational or the emotional – has produced the most satisfactory result in resolving conflicts?
2. Imagine that the knowledge leading to atomic energy had never been achieved. What are some of the things that might have been different about the period from World War II to today? Can you say whether the world would have been better or worse off?
3. Imagine an intelligent species evolved on a planet habitable like the Earth, but with a perpetually opaque atmosphere so that the stars could not be seen. How might they regard themselves and the nature of their world as a planet in such circumstances? Could they infer the presence of other stars, planets, and moons? Would there be any impetus to develop the ability to travel into space? Likewise, how would a species with the intelligence of humankind but restricted to the deep oceans define its "cosmos"?
4. A smaller and smaller fraction of the human species can see a star-filled sky at night, thanks to the increased amount of nighttime illumination used in cities. At the same time, an increased fraction of humankind has access to detailed images of the cosmos from large telescopes in space and on the ground. How do you think this shift in the nature of astronomical information will alter popular thinking about the cosmos in the next few decades? In the next few centuries?

General reading

Boorstein, D. J. 1983. *The Discoverers*. Vintage Books, New York.

Sagan, C. 1996. *The Demon-Haunted World: Science as a Candle in the Dark*. Ballantine Books, New York.

Reference

Snow, T. P. 1991. *The Dynamic Universe: An Introduction to Astronomy*, 4th edn. West Publishing, St. Paul, MN.

2

Largest and smallest scales

Introduction

In Chapter 1, we became acquainted with the scale of the solar system – the stage upon which planetary evolution is set. However, the formation of elements out of which planets and life came into being involved the universe of stars and galaxies – a scale much larger than the solar system – and the microscopic world of atoms, which involves size scales much smaller than that of our ordinary experiences. In this chapter we explore how cosmic distances are gauged, and then begin to acquaint ourselves with the basic building blocks of matter.

2.1 Scientific notation

Although the book is written with the nonmathematically inclined reader in mind, the discussion of numbers, both large and small, cannot be avoided if we are to gain a true understanding of Earth and its place in the cosmos. Numbers of interest in science range over enormous magnitudes (Figure 2.1). The number of protons contained in a single star, our Sun, is of order 1,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000; the size of an individual proton (itself made up of smaller elementary units) is of order 0.0000000000001 cm. (The term *of order* refers to how many powers of 10 a number contains, rather than the specific numerical value it has; hence 200 is of order 100, 40 is of order 10, etc.) These numbers are inconvenient to write down and manipulate in even the simplest mathematical expressions.

Hence *scientific notation* is universally used, where a number is expressed in terms of powers of 10. The number of protons in the Sun is of order 10^{57}; the size of a proton is of order 10^{-13}. To express the numerical value, in addition to the order of magnitude, one simply multiplies by the appropriate number. Hence, 5,000 is 5×10^3 and 0.004 is 4×10^{-3}. Any degree of precision can be handled readily; for example, 65,490 is 6.549×10^4 and 0.034256 is 3.4256×10^{-2}. Multiplication and division of such numbers is easy, the exponents in the power of 10 being added or subtracted, respectively, for the two operations.

The one drawback of scientific notation is that it dulls us to the enormous range of numbers that the scale of the universe demands. Somehow, writing 1.67×10^{-27} kilograms (the mass of a hydrogen atom) does not give us the same appreciation for the smallness of this number writing 0.00000000000000000000000000167 gives. (For US readers, one kilogram is roughly two and a quarter pounds.) In contemplating the history of planet Earth and its place in the cosmos, it is too easy to manipulate such numbers without first considering the philosophical implications of their gigantic or minuscule quality!

2.2 Motions of Earth in the cosmos

We view a universe continually in motion. The most obvious movements, apparent to even the casual observer, are the paths of the Sun across the sky on a daily basis and the rising and setting of the Moon on an apparently slightly less reliable basis. The equivalent nocturnal rhythm of the rising and setting of the constellations also is easily discernible, though much less familiar to increasingly urban populations.

Those who are more careful watchers of the sky will notice two longer rhythms, the march of a changing Moon progressively through the day and night skies on a 28- to 29-day basis, and the annual ritual of the slow climb of the Sun toward a more northerly path in the sky during summer and toward a more southerly path during winter (readers in the southern hemisphere should reverse north and south in the description). At any given location the Moon occasionally wanders into a region of darkness, and reddens in what is called a lunar eclipse. The Sun's light is partially blocked once every few years from a given location, and totally blocked much more rarely at any given place, in a solar eclipse.

Size (km)	Orders of magnitude larger than Earth	Significant distances
10^{23}	19	Diameter of universe
10^{22}	18	
10^{21}	17	
10^{20}	16	
10^{19}	15	Distance to nearest galaxy like ours
10^{18}	14	
10^{17}	13	Diameter of our galaxy
10^{16}	12	
10^{15}	11	
10^{14}	10	
10^{13}	9	Distance to nearest star (Alpha Centauri, Sun)
10^{12}	8	
10^{11}	7	
10^{10}	6	Diameter of solar system (Pluto, Sun)
10^{9}	5	
10^{8}	4	Astronomical unit (Earth, Sun)
10^{7}	3	
10^{6}	2	Diameter of sun
10^{5}	1	Diameter of jupiter
10^{4}	0	Diameter of earth
10^{-3}	7	Child
10^{-8}	12	Cell
10^{-13}	17	Atom
10^{-16}	20	Nucleus
10^{-20}	24	Electron

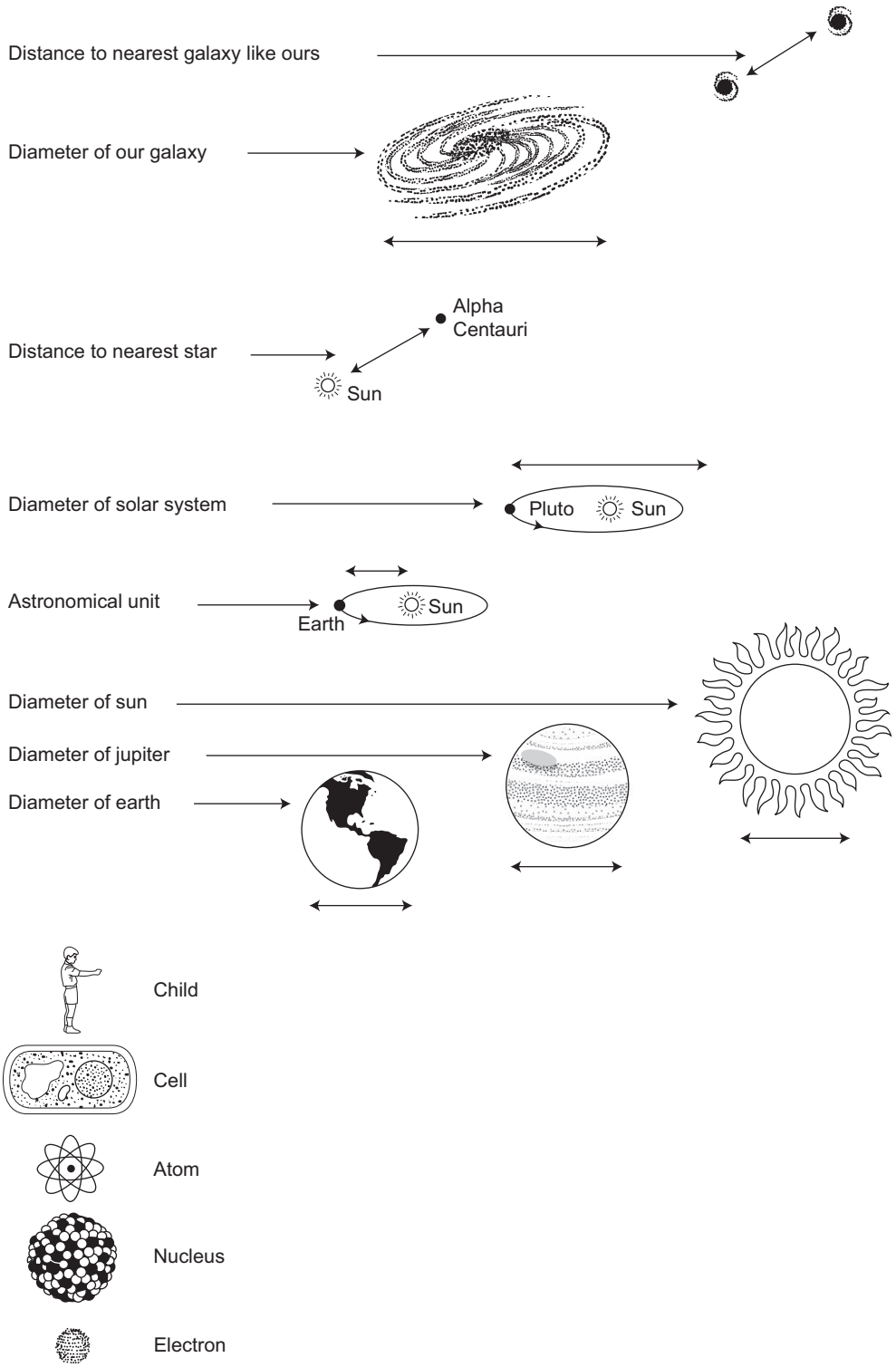


Figure 2.1 Sizes of various objects over the enormous range that the natural world encompasses. From Robbins and Jeffreys (1988) by permission of John Wiley and Sons.

Even more subtle motions are available in the skies for those with the patience to watch. Five "stars" in the sky can be seen, without a telescope, to move against the background of the fixed stars on paths that execute peculiar back-and-forth dances; the speed with which these planets (from the Greek *planetes*, meaning wandering) move varies greatly, corresponding to timescales of months to centuries to orbit the Sun.

All of these motions are fully understandable on the basis of the Copernican model of a spinning Earth, tipped modestly on its axis, orbiting about the Sun once each year, with other planets orbiting at greater or lesser distances from the Sun, and the Moon orbiting about the Earth. We take this picture, quite appropriately, as fact, but few of us have paused to ponder the subtleties associated with working out such motions. Furthermore, slight changes in the shape of Earth's orbit have affected climate on cycles of tens of thousands of years, and the presence of the Moon in orbit about Earth apparently has prevented rather extreme swings in Earth's axial tilt, which could have led to very large climate instabilities in the past. Far from being a quaint part of the traditional curriculum of science in schools, the arrangement of the Sun, Earth, Moon, and other planets is in fact critical to understanding the stability of, and variations in, our climate on a range of timescales.

We discuss such climatic issues in Part III, but now we return to the basics of Earth's motions through the cosmos. The perception of movement of the Sun and constellations through the sky is akin to our experiences as children (or adults) on a carousel, watching people, trees, structures, and so on swing past us in regular, repetitive cycles. Because there is little sense of acceleration on the larger, slower (and hence grander) carousels, very quickly one can experience the illusion of being on a fixed world around which the external "universe" is moving.

The Moon's motion is somewhat more complicated; because it is orbiting Earth once every 28 to 29 days, it rises and sets at significantly different times from one night to another. The analogy on our carousel is to watch a person who is walking briskly in the direction of the carousel's motion. Relative to fixed objects (standing adults, trees), our moving person will reappear later during each rotation of the carousel. Because our Moon is almost entirely illuminated by the distant Sun (some contribution from Earthlight is detectable on the otherwise unilluminated portion), different portions of the Moon are illuminated at different times of the month, creating *phases* (Figure 2.2).

The orbit of the Moon is not aligned with the apparent path that the Sun takes around our sky (called the ecliptic plane) but rather is inclined from it by about 5 degrees. Because of this, during the time of the month when Earth, the Sun, and the Moon are all aligned in a given direction (the times of full and new Moon), the Moon generally appears on the sky significantly above or below the path of the Sun. Only when the time of full Moon coincides with the Moon crossing the plane defined by Earth's orbit around the Sun – the *ecliptic* – do we have true alignment. At this time, the full Moon gives way to a *lunar eclipse*, in which Earth's shadow obscures the Moon, or the new Moon is replaced by the dramatic *solar eclipse*, in which the disk of the Moon blocks out the light of the Sun (Figure 2.2).

Eclipse prediction is not easy, because three motions are involved: the revolution of the Moon around Earth, the motion of Earth around the Sun, and the so-called *regression of nodes*, wherein the points at which the Moon crosses the plane of the Earth's orbit around the Sun rotate slowly in an 18.6-year cycle. This last motion can be visualized by imagining the orbit of the Moon as a circular glass sheet that cuts through Earth at a slight angle relative to the ecliptic. This sheet slowly revolves relative to Earth, completing one spin in 18.6 years. (The physical cause of the regression lies in the gravitational pull of the Sun, which exerts a torque because the lunar orbit is tilted or *inclined* relative to the plane of the Earth's orbit around the Sun, which is the ecliptic plane.)

These three motions are such that any particular sequence of eclipses recurs at an interval just over 18 years. The frequency of lunar eclipses is greater than the frequency of solar eclipses. Because Earth's shadow is much larger than the Moon when projected at the distance of the Moon from Earth, slight misses in crossing the node still produce a lunar eclipse. The lunar shadow is smaller and, coincidentally, the size of the Moon in the sky is just roughly that of the Sun. Thus the solar eclipse must occur very close to a node crossing for it to be total. Further, the orbit of the Moon around Earth is not a circle but an ellipse (see below); if the eclipse occurs when the Moon is farthest from Earth, the apparent size of the Moon is smaller than the Sun's disk, and a much less spectacular, *annular*, eclipse transpires.

Two remarkable cultures demonstrate both the subtlety and universality of tracking the rhythms of solar system objects. Stonehenge is a series of large rock monuments and circles laid out on the Salisbury Plain of England. The earliest such construction, most significant astronomically though least spectacular to the eye, is a large circle of 56 *Aubrey* holes, spanning roughly 50 meters across, with a so-called heelstone off to the northeast. This was set up by a Stone Age people about 4,800 years ago, perhaps a millennium before the large stone structures more familiar to tourists. Spurred by an initial suggestion by astronomer Gerald Hawkins, British astrophysicist Sir Fred Hoyle (1972) demonstrated that the 56 Aubrey holes could be used as an eclipse counter.

By moving stones representing the Sun and the Moon counterclockwise at certain prescribed rates (two holes every 13 days for the Sun and two holes each day for the Moon), one predicts the positions of the Sun and the Moon relative to the observer, on Earth, in the center of the ring. By moving two other stones, each 180 degrees apart, clockwise three holes each year to represent the precession of the lunar nodes, eclipses could be predicted reliably. When the Moon and the Sun are on opposite sides of the circle, and less than one or two Aubrey holes away from the node stones, a lunar eclipse would occur; when the Moon and Sun stones cross each other and are less than one or two Aubrey holes away from a node stone, a solar eclipse is predicted to occur (Figure 2.3). The counter scheme was not perfect, because about half of the predicted eclipses would not be visible in the skies above Stonehenge (the Aubrey circle representing the full 360 degrees of the sky including that beneath the horizon at Stonehenge); nonetheless, if correctly interpreted, it is a clever astronomical calculator.

Because none of the solar, lunar, or nodal cycles are exact multiples of the 56 holes, the counting rules are not exact. The marker positions would need resetting regularly by sighting the Sun and the Moon in the sky at key times of the year. The heelstone and nearby additional holes were used, according

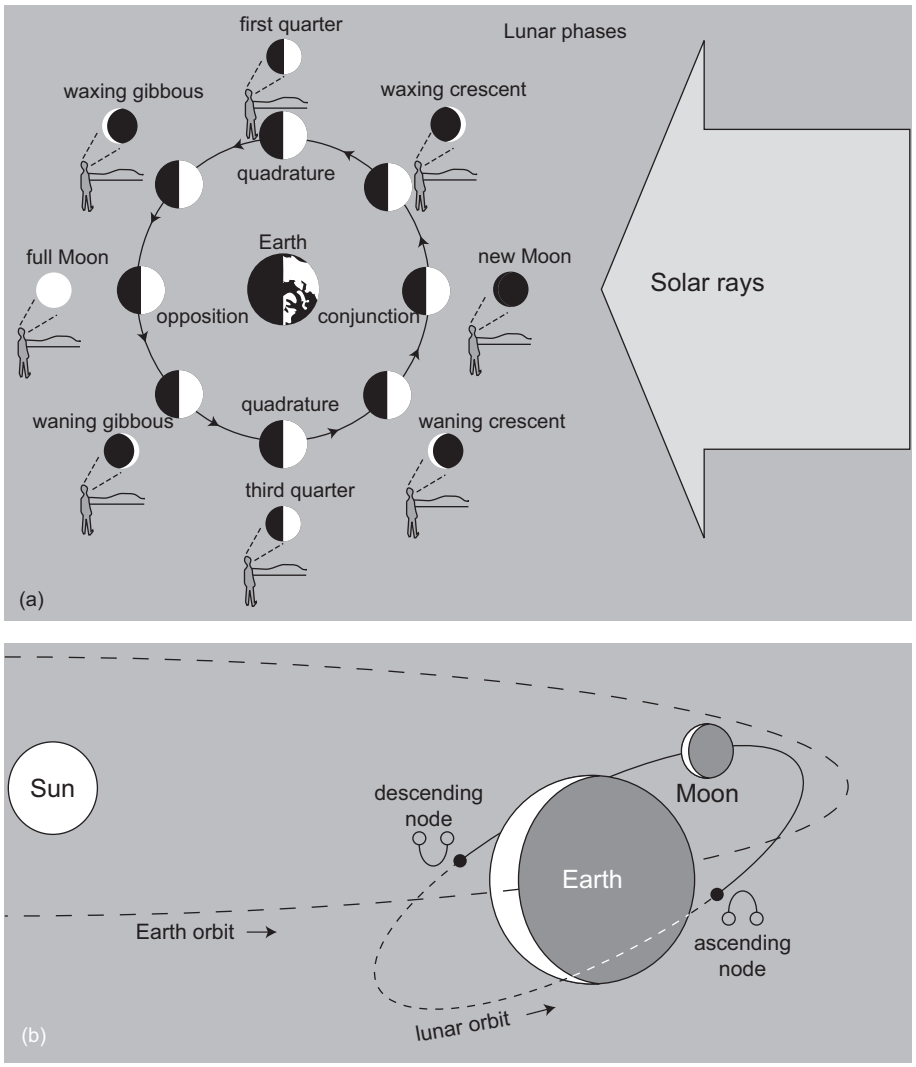


Figure 2.2 (a) Geometry of Earth, the Moon, and the Sun leading to the monthly cycle of phases; an Earth-bound observer's view is shown next to each corresponding lunar position (adapted from Snow [1991, p. 31]); (b) alignments of Earth, the Moon, and the Sun during total solar and lunar eclipses (after Hartmann [1983]).

to Hoyle's model, for sighting and hence correcting the board positions.

Intriguing as the eclipse counter itself is, Hoyle brought up the significant issue of what the node stones would have meant to the people of early Stonehenge. The need for node stones to determine when full or new Moon points would have eclipses must have been derived empirically, because as invisible mathematical constructs one cannot see nodes in the sky. Given that the Sun and the Moon are common objects of worship in many cultures – even our own, as technologically advanced as it is – it is interesting to ask what the Stonehenge people might have thought their node stones represented. It is tempting, as Hoyle wrote, to think that the node stones suggested to the Stonehenge culture the existence of a powerful yet unseen deity that controlled the motions of the Sun and the Moon. But this is piling speculation on top of an already interesting but speculative interpretation of an artifact, namely the eclipse counter itself!

The Mayan people live in the Yucatan peninsula region of Mexico and Central America. From roughly 100 B.C. to A.D. 900, they produced large numbers of stone sculptures, or stelae, on which a complex system of calendar dates was engraved. The classical culture of organized city-states had several calendars, including one of 365 days and a 260-day religious calendar. This latter is close to, but not quite, the orbit period of Venus. Astronomer–archeologist Edward Krupp (1983) also has suggested that it might refer to the interval between passages of the Sun across the high point (zenith) of the sky at the latitude of important Mayan cities, occurring in May and August. There are other astronomical and biological cycles of significance close to 260, including the human gestation interval.

Most striking about the classical Maya was their sophisticated numbering system for precisely recording dates of major events in their history. The system allowed for extension of dates back in time, and some Mayan sculptures do so – back to arbitrarily large values. The longest date recorded on a Mayan stela

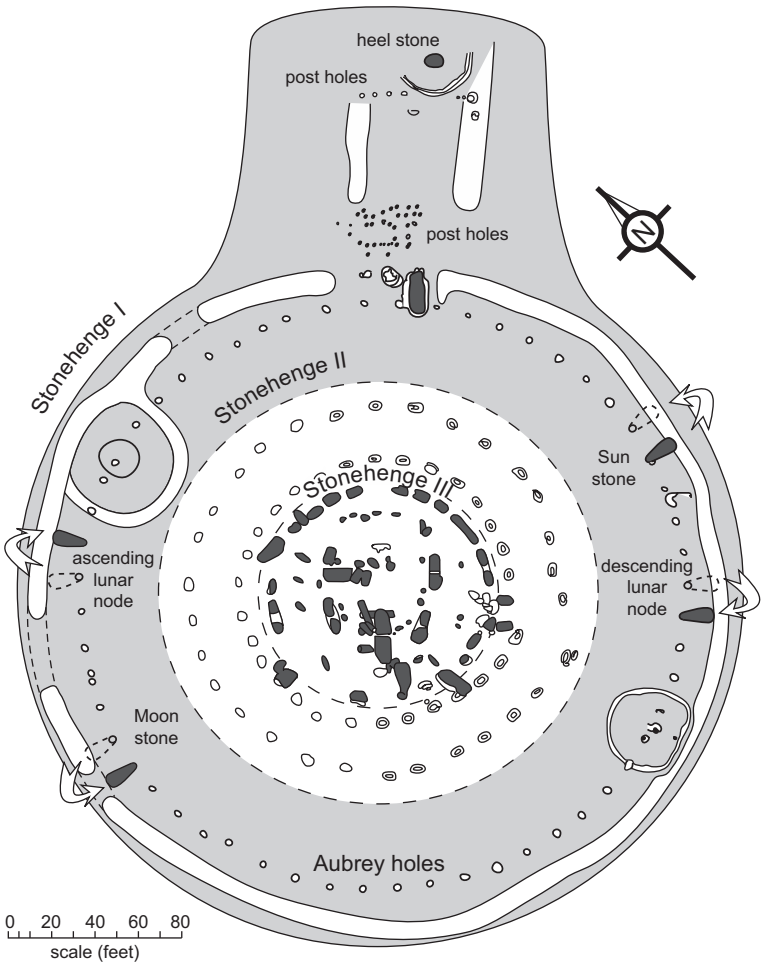


Figure 2.3 Map of the three stages of Stonehenge construction. The Aubrey holes and other sight points of Stonehenge I are identified. Adapted from Hoyle (1972 p. 22, Fig. 2.4) by permission of W. H. Freeman and Company.

corresponds to 1.4×10^{36} years, or 10^{26} times the age of the universe as determined by modern cosmology!

The classical Mayans regarded human history as one cycle embedded in nested sets of larger cycles. The Mayans established a "zero" date, prior to which events were played out by deities, which human events then mirrored. Hence history was already determined, in a sense, because it had been played out before on a larger scale. The progression of time was thus cyclical, but it was linear as well, in that the classical Mayan culture had a detailed chronological history of human events – battles, conquests, accessions – for which definite dates were assigned. Both significant human events and their mirrored supernatural events before the zero date often were pinned to particular points in the cycles of bodies in the sky, and the Mayans spent much time tracking and recording celestial movements so as to predict when significant events in human history might occur.

One might wonder whether this dual cyclical–linear concept of history arose out of the preoccupation of the classical Maya with calendar keeping, sky watching, and recording of dates, or vice versa. As with our own decimal system, where each digit placed to the left of preexisting digits represents a new power of 10 (and hence a larger supercycle of years, decades, centuries, millennia, etc.), the Mayan system of counting in twenties allowed cycles nested within cycles to be similarly expressed. In a different sense, our own Western concept of time also embodies both linear and cyclical elements; we will see this in our study of the history of Earth and its sister planets that forms the major part of the book.

2.3 Cosmic distances

2.3.1 The planets

Distances to the planets are precisely known today and spacecraft are sent to these bodies on a regular basis. But planetary distances began to be quantitatively determined only in the past few centuries. A German scientist, Johannes Kepler, around A.D. 1609 formulated a set of laws of planetary motion based on extensive observations by Tycho Brahe, a Danish astronomer. Kepler proposed that the planets moved around the Sun in *elliptical* orbits (Figure 2.4), that a given orbit swept out equal areas in equal amounts of time, and that the square of the period of the orbit was proportional to the cube of the planet's mean distance from the Sun. (No understanding of *why* the planets obeyed these laws came out of Kepler's proposition, at least not

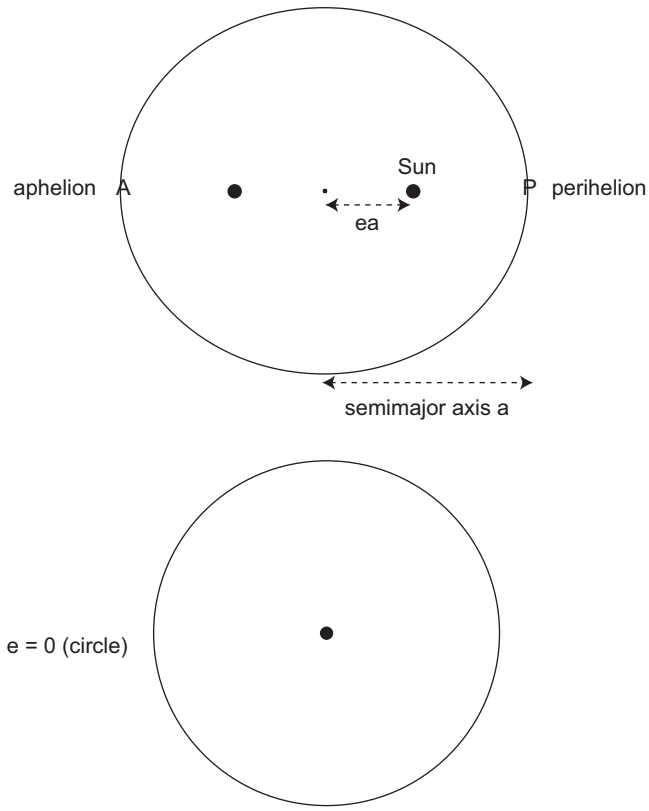


Figure 2.4 Planetary orbits are ellipses, with two "foci", at one of which is the Sun. The shortest (perihelion) and longest (aphelion) distances to the Sun along the orbit are labeled. The distance from the small point at the center of the ellipse to either P or A is the semimajor axis a; the distance from one focus to the central point is the semimajor axis multiplied by the eccentricity of the orbit, e. When e = 0 the two foci are coincident, and a circular orbit results.

immediately; the English mathematician and physicist Sir Isaac Newton decades later postulated the existence of an attractive force associated with the mass of an object, namely gravity. Kepler formulated his laws solely to fit observations; this is an excellent example of an *empirical* model.)

Given Kepler's laws and knowledge of the distance of Earth from the Sun, one can work out the distances to the other planets merely by determining the time it takes each to rotate once around the Sun, that is, the *period* of the orbit. Earth orbits the Sun once in one year. Jupiter orbits the Sun once in just under 12 years; taking the ratio of these periods, squaring it, and taking the cube root yields a mean distance from the Sun for Jupiter of about $5\frac{1}{4}$ times the Earth–Sun distance, or 5.25 AU. Pluto, the most distant planet, has an orbital period of 249 years and hence a mean distance of 39 AU. (We have seen only one-fifth of its orbital path around the Sun, but it was possible to fit an ellipse to its path and hence determine a period very soon after its discovery in 1930.)

However, there is one missing link: the distance of Earth from the Sun. We have expressed planetary distances in terms of Earth–Sun distance, but this is not very satisfying. The distance from Earth to the Sun was tackled by the Greek scientist Aristarchus of Samos, who worked out that when the Moon was exactly half full, the Sun–Moon–Earth would make a right triangle. The angle between the Moon and the Sun in the sky viewd by an observer from Earth then yields, by simple trigonometry, the Earth–Sun distance, provided one knows the Earth–Moon distance. This distance, in turn, was found by comparing the size of the Moon to the size of Earth's shadow projected against the sky (and revealing itself during a lunar eclipse), yielding a size for the Moon roughly one third that of Earth. This then led to the lunar distance from Earth, and hence the Earth–Sun distance. Unfortunately, Aristarchus was unable to accurately measure the Moon–Sun angle in the sky, and did not get the right answer, but conceptually this is one valid procedure for getting the Earth–Sun distance, about 150 million km.

A second, ultimately more precise, determination of the scale of the solar system was obtained by observing the *parallax* motion of the planets. This technique is fundamentally important for determining distances to the nearby stars, beyond our solar system, and so we explain it in that context, it the next section.

2.3.2 Nearby stars and planets redux

No stars that we see in the sky orbit the Sun. Instead, the Sun is one of 100 billion stars that orbit about a common center of gravity; this enormous collection of stars is called the Milky Way Galaxy.

To measure the distance to stars relatively near our solar system, the optical effect of parallax can be used. Parallax can be easily experienced by holding a pencil in front of your eyes and alternately closing your left and right eye. The pencil is seen to shift against the background. The same effect is present when stars closest to us seem to shift the most during Earth's annual orbit around the Sun. The 300-million-km diameter of Earth's orbit serves as the equivalent of the separation between your eyes in the pencil experiment. By measuring how much stars shift against the background during observation (with highly sensitive telescopes) six months apart, absolute stellar distances are obtained. The nearest star, the Alpha Centauri multiple star system, is four light-years away. (A light-year is the distance light travels in a single year, about 10^{13} km.) Beyond a few thousand light-years from Earth, parallax shifts are too small to be measured and other distance techniques must be used. However, a satellite named *Gaia* to be launched in 2013 will measure distances so precisely that this scale may be extended out to tens of thousands of light years.

The parallax technique itself has led to common use of a unit of stellar distance different from the light year: The *parsec* (from parallax-second) is the distance to an object that exhibits a parallax shift of 1 arc-second in the sky, which is 1/3,600 of a degree of angle, the full sky being 360 degrees around. Defined as it is for a baseline corresponding to the diameter of Earth's orbit, the parsec works out to be 3.26 light-years.

Aristarchus' model of an Earth moving around the Sun was disputed by other Greeks because they could not see relative shifts in the position of stars from one side of Earth's path around the Sun to the other. We know now that the problem lay in the great distance to even the nearest star, which results in a parallax shift too small to be detected by the Greeks, who had no telescopes. The planets of our own solar system exhibit a larger parallax, but even this is difficult to see because the

sizes of the planetary disks themselves obscure the parallax shift. When the planet Venus transits (passes across the disk of) the Sun, observations from different parts of the Earth can made of the precise times when the disk of Venus enters and then exits the bright disk of the Sun. In this way the parallax may be determined and hence the value of the astronomical unit. In practice turbulence from our own atmosphere limits the accuracy of the observation. It required data from four transits in the eighteenth and nineteenth century to obtain a value of the astronomical unit close to the precise one known today – a value finally obtained from bouncing radio signals off of the surface of Venus, timing their return to the Earth, and using the fact that the speed of light in vacuum is a known constant.

2.3.3 Nearby galaxies

Beyond the distances accessible to parallax measurements one must use indirect techniques. If all stars were the same brightness, we could measure distances by comparing a star's brightness with that of one whose distance has been determined, for example, by parallax observations. The technique would be akin to looking out over a city from a hilltop and gauging distances to various streets by the apparent brightness of their streetlights. Light spreads out and dilutes in two dimensions as it moves away from its source, so that the apparent brightness of an object must decrease as the square of the distance from it. Precise measurement of the brightness, then, is a unique measure of distance as long as the intrinsic brightness is known, and there is no absorption of light by dust or gas between the observer and the source. In the case of our streetlight analogy, several effects could create an error: some streetlights have lamps that are intrinsically brighter than others, because of both the type of lamp and its time in service.

Stars as well vary greatly in their intrinsic brightness, depending primarily upon their age and mass (amount of material they contain). The brightness range for long-lived stars, so-called main sequence stars (see below), is 10 orders of magnitude; for stars in various early and late stages in which dynamic processes are occurring, the range can be much larger. Thus stellar brightness generally is not a useful measure of distance.

Luckily, there exists a group of stars whose intrinsic brightness is related to another property that can be measured independent of the star's distance from us. These are the so-called *Cepheid*-variable stars, which pulsate in brightness in a rhythmic way. The more rapidly a particular Cepheid pulsates, the dimmer it is. The relationship has been determined empirically for Cepheids that are close enough to Earth for their distances to be measured by parallax, and hence for the star's intrinsic brightness to be worked out. The faster–dimmer relationship seems to hold so well that the intrinsic brightness of any given Cepheid is predictable from the pulsation period.

Cepheid pulsation periods can be measured out to great distances, limited only by the ability to detect the pulsations in very faint sources (faint because of the great distance). From the pulsation period, the star's intrinsic brightness thus can be determined. With large ground-based telescopes, the technique has been extended out to the neighboring galaxies, millions to tens of millions of light-years distant. The Hubble Space Telescope, positioned above Earth's distorting atmosphere, has been used to observe Cepheids as far away as 100 million light-years. The extent of our own Milky Way Galaxy is determined from this technique to be of order 100,000 light-years across.

2.3.4 Beyond the galactic neighborhood

For more distant galaxies, Cepheids are too faint to be detectable and hence to have their pulsation periods measured. Distance determinations in the absence of Cepheid detections are much less precise. Certain stellar explosions, called *Type 1A supernovas*, seem to produce a characteristic peak brightness as the star explodes and then dims. By observing such supernovas in nearby galaxies for which Cepheid variables are measurable (to determine the galaxy's distance), the Type 1A supernova brightness can be calibrated. Because such explosions are enormously bright, millions of times that of a Cepheid variable, they allow the distance scale to be extended outward to several billions of light-years, a significant fraction of the size of the known universe (see the next section).

For galaxies in which no serendipitous supernova explosion is observed, the brightness of the whole galaxy must be used as a distance indicator, at least out to several hundred million light years. One might wonder whether this is a reliable technique, given the wide variation in the brightness of different stars. However, various tricks can be used that take advantage of observations of closer-in galaxies. Spiral galaxies, so named because the stars trace out a distinctive spiral shape as they orbit a common center, are particularly important in this regard. The more stars present, the more massive the galaxy, and the faster will be the rotation of stars around a common center. But the more stars present the brighter the galaxy will be overall. Hence the so-called Tully–Fisher relation allows the intrinsic brightness of a galaxy to be estimated from the rate of rotation using nearby galaxies to establish the rule. What is required is a way to remotely measure the velocity of the stars as they rotate around the center of a given galaxy. This, in turn, comes from observing the change the velocity of a luminous object has on its color, an effect we describe in the next section in the context of the final rung of the distance ladder, the measure of the size of the known universe.

2.3.5 To the farthest edge of the universe

Hearing the horn of a passing car is an odd experience, if you remember that most car horns are designed to produce a sound of a single pitch. The pitch of a passing car horn is higher when the car is approaching and lower when it is receding. This phenomenon is known as the *Doppler shift*, and it applies equally to waves of light and to sound. Because light, like sound, travels at a finite speed, the relative motion between source and observer causes waves to bunch in the oncoming direction and to be stretched out in the receding direction.

We discuss the nature of light in Chapter 3, but for now it suffices to construct a mental picture of light as the movement of waves of electric and magnetic, or *electromagnetic*, energy through space. The distance between each crest of the wave determines the color of light as perceived by the eye or measured more precisely with an instrument called a *spectrometer*. (This is something of an oversimplification; light emitted by

natural sources typically consists not of a single wavelength but a combination of many wavelengths, which, overall, yields the perceived color of the light.) An observer moving toward a source of light will perceive the waves to be bunched, and hence the color of the light shifted to the blue. An observer moving away from the source will see a shift to the red in the color of the light. Because of the enormous (but finite) speed of light, 3×10^5 km per second (a billion kilometers per hour), blue and red shifts are not noticeable at speeds with which we are familiar.

Spectrometers can measure the color of galaxies very precisely. It has been found that more distant galaxies appear to be redder. There are three possible causes of the reddening: increasing amounts of dust absorbing blue light between the observer and the galaxy, very strong gravitational fields near the galaxy, or a high recessional velocity leading to Doppler shift.

The first possible cause is eliminated by measuring the positions of discrete lines in the spectrum (Chapter 3); these are shifted toward the red by the Doppler or gravitational effects, but are unaffected by intervening dust. Gravitational fields as agents of red shift are a bit harder to eliminate, and may occur in some cases. However, in general, astronomers do not see other phenomena thought to be associated with strong gravitational fields when looking at most distant galaxies, and hence the bulk of galactic reddening should not be caused by strong gravity.

The third explanation seems to be the simplest and is supported by direct distance measurements to galaxies that are relatively near. It was American astronomer Edwin Hubble who first came to the sobering conclusion some half-century ago: the more distant the galaxy, the faster it is receding from us. In effect, the universe is flying apart from itself as if born in an enormous explosion.

The velocity–distance relationship can be established using the cosmic distance-measurement techniques described above, and then can be extrapolated beyond those techniques to determine the distance to the farthest galaxies based on their red shifts (hence their recessional velocities). Because the relationship between red shift and distance must be calibrated using nearby galaxies and more direct distance measures, it is sensitive to errors in the calibration techniques. A more precise determination of the overall size of the cosmos, and hence its age, comes from the echo of an enormous and ancient explosion heard in radio static, which permeates all directions of the sky and which can be detected on sensitive radio telescopes. This cosmic background radiation marks the horizon on the sky beyond which, during the very young universe, space was so hot and cluttered with dense matter that electromagnetic energy could not move freely through it.

As the universe has expanded over billions of years, this horizon has receded; it is now so distant that the original energy of the explosion is red shifted into the radio part of the electromagnetic spectrum, which is defined and described in Chapter 3. Mapping of this background radiation by a sensitive orbiting satellite experiment, *Cosmic Background Explorer*, indicates a remarkable uniformity that tightly constrains detailed models of how the cosmic explosion, or *Big Bang*, actually proceeded. However, the background is not completely uniform, and measurements of subtle variations in the intensity of the background radiation in different parts of the sky made by another Earth orbiting satellite (the Wilson Microwave Anisotropy Probe) more tightly constrain the time since the explosion to 13.7 billion years. This is when the expansion of the universe began. But it cannot be the simple expansion of matter into a static void. The observed fact that on ever-larger scales the universe seems to be moving away from us, in all directions, might lead one to conclude that we are at the center of the cosmos, an unpalatable notion in view of the fact that nothing else about the Earth's place in the cosmos seems "special". In fact, it is space itself that is expanding, and this means that the appearance that all is "flying away from us" does not imply we are in a special place in the cosmos: all observers everywhere see the same effect. In consequence, the red shift of the galaxies cannot be a Doppler shift in the strict sense, since such a phenomenon is the result of movement through a fixed medium. The galactic red shift is better thought of as a signature of the expansion of space itself, a phenomenon with no direct analogue in our daily existence. (Therefore, our discussion above of the red shift of spectra of more distant galaxies was imprecise, but necessarily so given that we cannot dwell too long on cosmology before turning to the main subject of the book.)

As creatures of a three-dimensional reality, we can visualize the expansion of space itself only by making the thought-experiment of reducing the number of dimensions by one. The analogue to the cosmos, then, is the *surface* of a balloon in the process of being inflated. Imagine space to be that two-dimension surface, within which are embedded truly two-dimensional observers that are fixed on specific points of the ballon surface, and therefore move apart from each other as the surface area of the balloon grows. (If you wish, you can imagine tiny bugs crawling on the balloon surface, even though they are not truly two-dimensional.) The particular geometry of the balloon surface, that of a closed sphere, is almost certainly not a good two-dimensional analogue of the shape of the cosmos in three dimensions, but it illustrates one important effect: to an observer anywhere on the balloon, all other observers seem to be moving away from him or her. Every point on the surface of the balloon seems special in terms of being the "center" of the expansion, but there is in fact no center – no place on the surface of the balloon is special. Observing the expansion of the cosmos "around us", the impression of a center is an illusion caused by space itself expanding. We can only know about the region of space within which light has traveled since the beginning – the Big Bang. But space could extend beyond this so-called "horizon", and have vastly different properties there. It need not even have the same number of spatial dimensions – as pointed out so eloquently by Harvard physicist Lisa Randall in her remarkable *Warped Passages* tour of hyperdimensional space and time. (What space is expanding into, and what initiated the expansion, are deeply fascinating problems in and of themselves.)

The expansion of space itself has another important consequence, namely that while matter is constrained to move at a velocity less than that of light, the expansion of space is not. The initial expansion of the cosmos must have included a very brief phase in which the scale of everything suddenly increased dramatically, called "inflation" (but not to be confused with the gradual inflation of our balloon analogous to the long-term expansion of space itself) – a scale change required to explain the relative uniformity of the distribution of galactic clusters on the sky (Figure 2.5). But even the subsequent expansion of

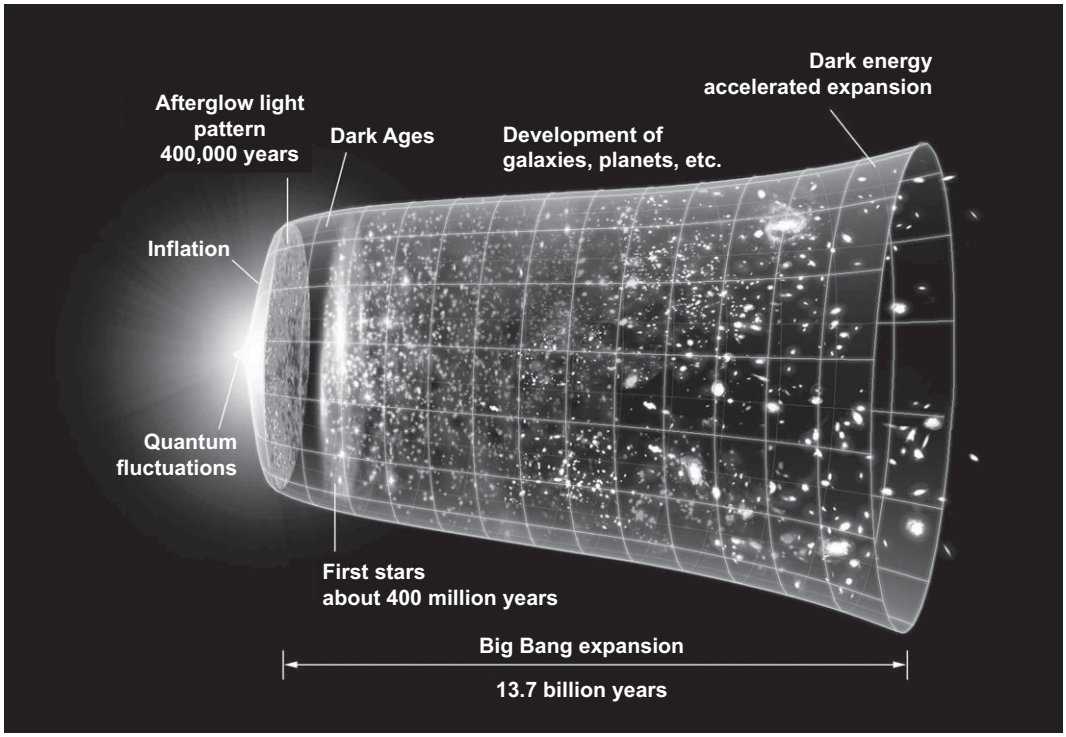


Figure 2.5 Schematic history of the entire cosmos, in which time flows from left to right. Immediately after the Big Bang all of the cosmos is dominated by fluctuations on a quantum scale, and coherent macroscopic reality as we know it does not exist. Then the scale of the universe greatly expands, in a phenomenon known as inflation, leading to the afterglow light pattern of fluctuations in the "cosmic microwave background" radiation that we see today. As the first stars form, some 400 million years after the Big Bang, formation of elements (Chapter 4) begins. Expansion of the cosmos appears to be under acceleration today, associated with a repulsive "dark energy" whose nature is not understood. Figure courtesy NASA WMAP Science Team. See color version in plates section.

space has not been uniform: Hubble Space Telescope and other observations reveal that the expansion of the universe is accelerating, and began doing so about halfway through cosmic history. Here again, it is not that the clusters of galaxies (the "islands" of matter in the cosmos) are accelerating away from each other in a fixed void. It is the space within which they are embedded that is accelerating. Returning to our two-dimensional analogy, the balloon is being inflated at an ever-increasing pace. This requires a scale effect intrinsic to the geometry of space itself (Einstein's "cosmological constant"), or is indicative of a hitherto undetected repulsive force or negative pressure associated with some unknown energy in the vacuum of space itself . The latter interpretation has gained the most favor among cosmologists, and this form of inferred-but-not-understood energy is called "dark energy".

To add to the exotic nature of reality on large scales, even the simple spin of the spiral galaxies – how fast the inner and outer parts rotate around their common center – does not seem to follow the law of gravity (Chapter 3) formulated by observing the motion of the planets around the Sun. Because there is no other overt violation of this law, it is most straightforward to postulate a form of matter – "dark matter" – that cannot be seen but is present in sufficient abundance in galaxies to exert an additional gravitational pull. Intriguingly, to explain the rotation of material in galaxies requires that dark matter be five or six times more abundant in the cosmos than ordinary matter. As we discuss in Chapter 4, matter can be converted to energy, and were one to convert all the ordinary matter and dark matter into energy, these would constitute at present only a quarter of the total energy of the cosmos: the remainder is dark energy if the cosmic acceleration has been correctly measured and interpreted. It would seem that most of the universe is unobservable, exotic, or both.

Dark matter appears to be exotic matter that does not interact with light, but its identity remains elusive. There may be several contributors to dark matter, among which are neutrinos, subatomic particles that have been detected, but are ghostly in that they interact only weakly with normal matter (and hence, for example, pass through solid rock). To understand the Earth and its history, we must understand matter at the atomic and subatomic scale, and it is to this subject we next turn.

2.4 Microscopic constitution of matter

All forms of matter with which we have direct familiarity are composed of a relatively small number of *chemical elements*. Of these, 111 such elements are known, of which roughly 90 occur in nature. The rest have been made in the laboratory; although some of these may occur in nature under extreme conditions (supernova explosions), they are too short-lived to be detectable.

Elements occur as chemically irreducible bits of matter called *atoms*; these are the smallest particles of matter that retain the chemical identity associated with elements. In our own lives we mostly encounter atoms combined into composites called *molecules*. Inside stars like the Sun, temperatures are high enough that atoms themselves are partly broken apart into negatively and positively charged pieces; the resulting form of matter is called a *plasma*. Very compact dense objects such as neutron stars (the collapsed remains of massive stars) contain matter under such extreme pressure that only subatomic particles called *neutrons* can exist.

The search to understand the essence of matter, specifically whether it could be infinitely divisible or was reducible only to some definite elemental particle, began (in documented history) with the ancient Greeks. Democritus was a fifth-century B.C. Greek philosopher whose preference for an atomic model came largely from his views on human progress. If the material of the universe was built up of elementary particles, then the possibility existed that it was finite in complexity and hence understandable. The Roman poet Lucretius, in the first century B.C., elaborated on the philosophical aspects of atomism, arguing that, if atoms obeyed a set of natural laws, then everything in the universe obeyed such laws and the supernatural did not exist.

In medieval times in Europe, the concept that the universe might be understandable as a finite collection of fundamental particles clashed with theological views, but in any event could not be tested and elaborated until experimental-based laboratory science developed in the seventeenth and eighteenth centuries. Prior to that, the unsuccessful attempts by the "alchemists" to transform common metals such as lead into precious ones such as gold did little to advance understanding of the nature of matter. Ironically, however, these endeavors philosophically foreshadowed the discovery of nuclear processes, although they fell hopelessly short of the energies required to the alchemists.

Laboratory evidence for a small number of different types of elements as the fundamental constituents of matter began to accumulate in the eighteenth century. Many common materials could be shown always to consist of irreducible proportions of other substances. Furthermore, when more than one sort of compound could be formed out of two elements, the ratio of the amounts of a particular element in one compound compared to the other could be expressed as small whole numbers. For example, the amount of oxygen in carbon dioxide is just twice that in carbon monoxide, for a fixed amount of carbon. These and other observations led Lavoisier in the eighteenth century, Dalton in the early nineteenth century, and others toward an understanding that the world was indeed composed of a small number of elemental building blocks.

Experiments in the late nineteenth century involving electrical discharges in gases began to elucidate the nature of atoms as being composed of negatively charged *electrons* and positively charged *protons*. Experimentally, it was found that opposite charges (plus and minus) attract each other, whereas like charges repel. To ensure electrical neutrality, it was thought initially that these must be mixed uniformly in the atom. A very different distribution of these charges was revealed by Ernest Rutherford's famous experiment in the early twentieth century. Rutherford fired a narrow beam of α *particles*, positively charged fragments of atoms, at a very thin (4×10^{-5} cm) foil of gold.

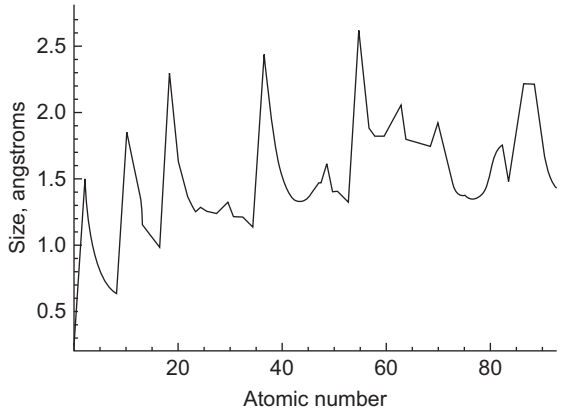


Figure 2.6 Size of elements as a function of their atomic number.

He then measured the various directions of scatter of the α particles, which are repulsed by the positively charged component of the atom. Most of the α particles were not deflected, but those that were scattered either nearly directly back or through very large angles. The results required that the positive charge of the gold atoms be concentrated in a very small volume, the *nucleus*, relative to the total volume of the atom, which is balanced by the negative charge of an equal number of electrons occupying a much larger volume.

It previously had been determined that the negatively charged electrons carried very little of the mass of the atom, and hence both the mass and the positive charge of the atom must reside in the very small nuclear space, worked out from experiment to be 10^{-12} of the volume of the atom itself. Furthermore, although it was found that the heavier elements had more protons in the nucleus, and correspondingly more electrons to ensure electrical neutrality, the mass of the elements did not increase linearly with the positive charge of the nucleus. The neutron, with zero electric charge, was postulated and discovered in the early 1930s. The proton and neutron have nearly the same mass, about 1.7×10^{-24} grams, and roughly 1,800 times the mass of the electron.

Elements were found to be defined by the number of protons, referred to as the *atomic number*. Atoms range in size, defined by the distance from the center of the nucleus to the outmost electron, from roughly 0.3 to slightly over 2.6 *angstroms*, where an angstrom is 10^{-8} cm. However, the elements do not increase linearly in size with increasing atomic number (Figure 2.6). Instead, the atomic size zigzags in a fashion that is correlated with the chemical properties of the elements, or more specifically, the particular manner in which elements will bond with each other to make the enormous variety of materials in the world around us.

Patterns in chemical properties of the elements were recognized in the eighteenth century by French chemist Antoine Lavoisier. In the late nineteenth century the Russian scientist Dmitri Mendeleev constructed a so-called *periodic table* of the 60 or so elements known at the time, based on their experimental chemical properties. The modern version of this table, shown in Figure 2.7, encapsulates the essential characteristics of the different types of atoms in the way they bond. The utility of the table was demonstrated repeatedly in the nineteenth century as

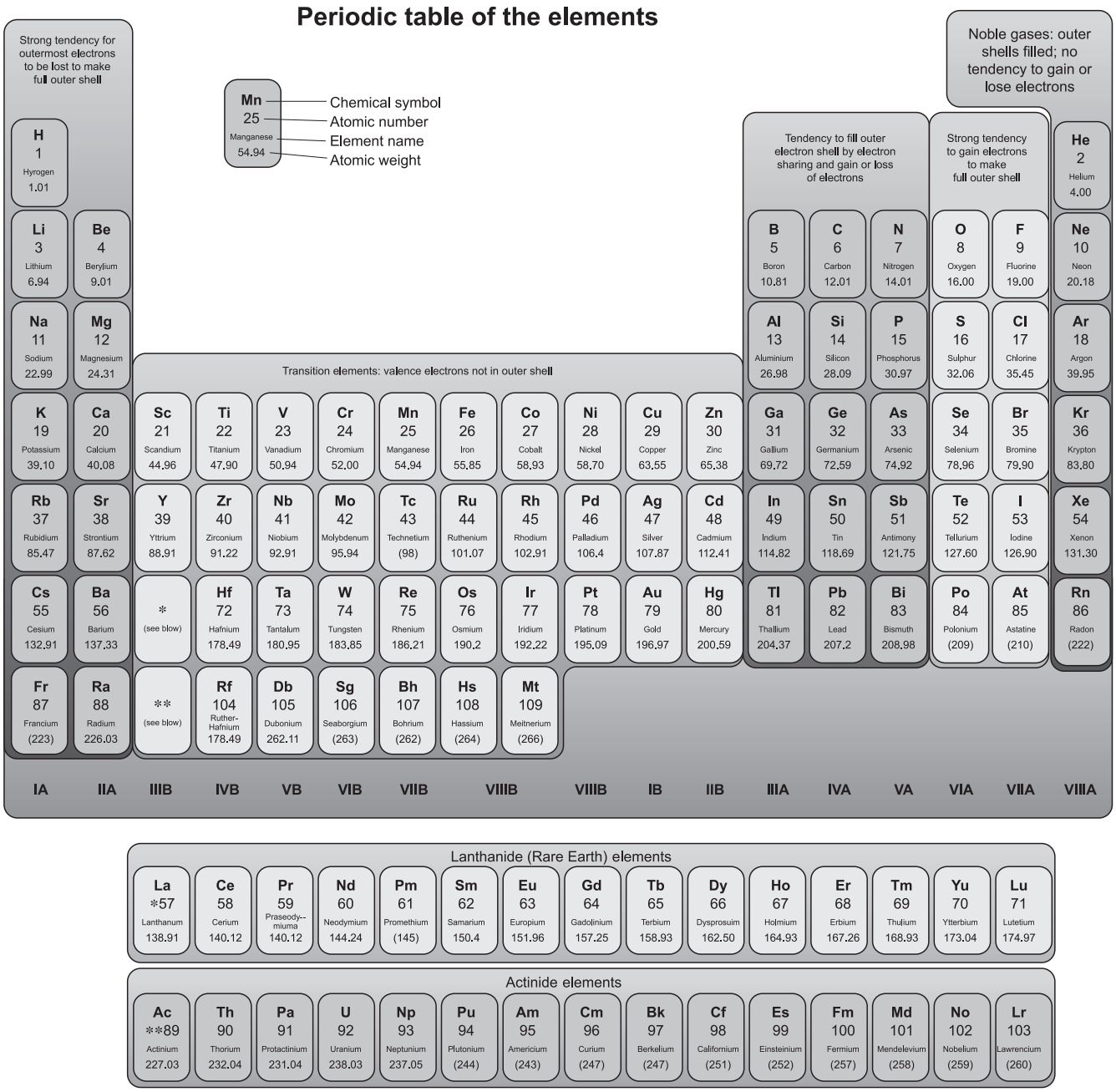


Figure 2.7 Periodic table of the elements. The elements are arrayed according to their bonding tendencies, as described along the top of the table. The term *valence* refers to the electrons, usually the outermost ones in the atom, that most actively participate in the joining of elements to make molecules. The symbol, full name, atomic number, and atomic weight are given for each element. The atomic weight, by convention, is the average value for the stable (unchanging) isotopes of each element, weighted by abundance as described in the text. Some of the heavier elements have no stable isotopes, and so, approximate atomic weights for the longest-lived forms of these elements are given in parentheses. The Lanthanide and the Actinide elements represent series of chemically almost identical elements. Each of the series occupies a single place in the table as indicated by the asterisks. Names of the six heaviest elements are those officially adopted by the International Union of Pure and Applied Chemistry in September 1997.

properties of as yet undiscovered elements were predicted from the table.

Elements are arranged in the table in order of increasing atomic number, but in a periodic fashion such that elements that lie in the same column have similar chemical bonding properties. For example, hydrogen, lithium, sodium, and the other *alkali metals* down column IA all have a strong tendency to form two-atom (*diatomic*) molecules with fluorine, chlorine, bromine, and other *halogens* (or salt-formers) in column VIIA. Typical compounds are HF, HCl, NaCl, LiF, LiCl, etc. (chemical symbols for the elements are given in the periodic table). Two atoms in column IA will combine with one atom of the

elements in VIA, the *chalcogens* (or ore-formers), so that H_2O, H_2S, Na_2S, and K_2S are all common *triatomic* compounds. The elements in column VIIIA, the *noble gases*, do not chemically bond or do so only weakly under relatively extreme physical conditions.

Note that, because a given element can bond with many other elements from different columns, there is a large degree of complexity in the number of molecules that can be formed. Furthermore, elements toward the middle of the table exhibit a tendency to combine in many different ways, even with a given second element, for example, carbon monoxide (CO) and carbon dioxide (CO_2). The origin of bonding patterns, and hence of the periodic table, lies in the particular number and configuration of electrons that an element possesses. Recalling that an element is defined by its atomic number, or number of protons, this also must be the number of electrons the element possesses to remain electrically neutral. Because the electrons move around in a volume that is much larger than the volume of the nucleus, it is logical that the interactions between the electrons determine the bonding between atoms.

An understanding of how electronic structure arises came primarily through the development of *quantum mechanics* during the early to mid twentieth century. Quantum mechanics is a branch of physics that deals with the behavior of matter at very small spatial scales. Much of this understanding came through studying the light released or absorbed by electrons in the atom, a subject we take up in Chapter 3. The key concept is that electrons possess definite values of energy as they move around the nucleus of the atom. The lowest energy level lies closest to the atom. Increasing energy levels are defined in terms of the pattern of electronic motion around the nucleus at a given energy level. Electrons have the property that they cannot exist identically in the same energy level with another electron. Two electrons can occupy one energy level only if a certain intrinsic property, called spin, is oppositely directed in the two electrons.

Certain preferred numbers of electrons exist at different energy levels, but most elements either have a deficit of electrons relative to the preferred number or have one or more excess electrons. There is a tendency then for elements to bond with each other in such a way as to produce the "right" number of electrons in each energy level. Direct transfer of electrons (*ionic bonding*) may occur, or the elements may simply associate closely in space so as to share electrons (*covalent bonding*). The different columns in the periodic table group the elements that have a certain excess or deficit of electrons relative to the preferred number for given energy levels. The table thus is a guide to how different elements are likely to bond. Elements on the left tend to donate electrons; those on the right need to acquire electrons. The rightmost column consists of elements that have the preferred number of electrons in all energy levels; these elements are chemically nonreactive and are called *noble gases.*

Elements in the middle columns of the table can either donate or acquire electrons with nearly equal likelihood. This leads to the many kinds of chemical bonds between these elements and others. Carbon, for example, can bond in many different ways with other elements. It is this versatility that is part of the reason why carbon is the most ubiquitous element in biological systems, playing a number of crucial roles. Overall, the variety of different propensities for bonding among the elements leads to the rich diversity of material properties in the universe (Figure 2.8).

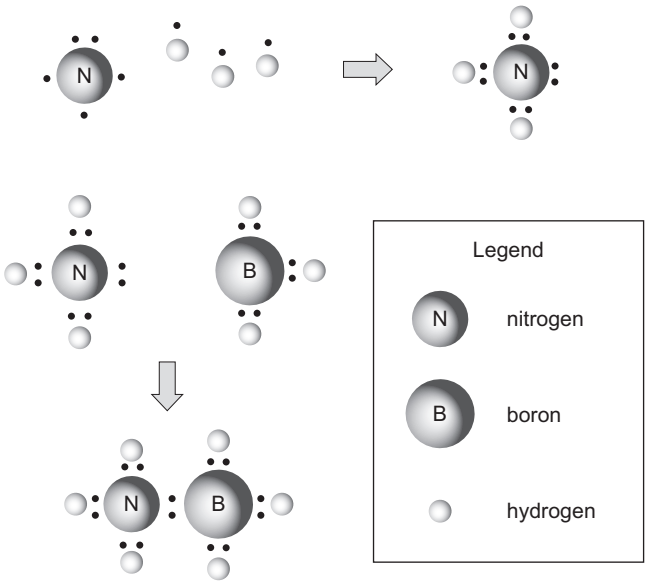


Figure 2.8 Simple example of how elements transfer or share electrons to achieve long-lived states: bonding of nitrogen and hydrogen to form ammonia, and bonding of ammonia and borane. Electrons that may be shared are indicated by dots; other electrons (not shown) are in energy levels much more tightly bound to the nucleus and do not participate in a significant way.

The neutral particles in the nucleus, the neutrons, do not affect the chemical properties of an element in a primary way. However, the same element can possess different numbers of neutrons, and these different varieties of the same element exhibit modestly different chemical properties. The total number of protons plus neutrons in a given atom is called its *atomic weight*. Atoms of the same element that have different atomic weights are called *isotopes* of that element. The average atomic weight of a given element in nature is listed in the periodic table of Figure 2.7. The fact that this is a fractional number reflects the mix of different isotopes in a natural sample of that element.

Considering hydrogen as an example, the primary isotope, sometimes called *protium*, has no neutrons and one proton, for an atomic weight of 1 (the small mass of the electron, by convention, is not included). The next isotope of hydrogen, called *deuterium*, has one neutron and one proton for an atomic weight of 2. Tritium is next, with two neutrons and an atomic weight of 3. Taking the abundances of the three isotopes found commonly on Earth yields the average atomic weight for hydrogen given in Figure 2.7 (1.00797). Remember, however, that *each individual atom has an integral atomic weight*, the mass of the electron not being included. No other elements have separate names reserved for their different isotopes. Instead, the atomic weight is attached as a superscript, so that protium (hereinafter referred to as hydrogen), deuterium, and tritium are ^1H, ^2H, and ^3H. The presence of

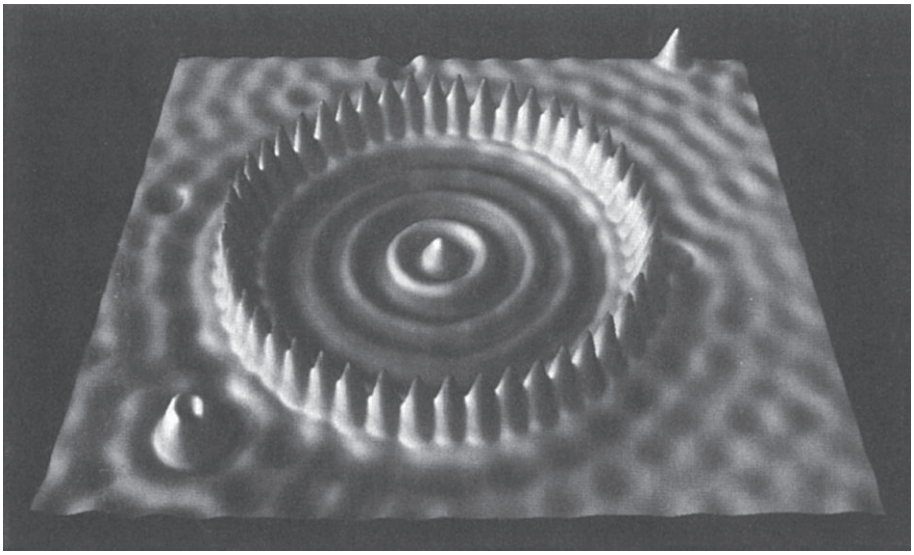


Figure 2.9 Image, using scanning tunneling microscopy, of an electron trapped in a ring of iron atoms. The electron is not merely a particle confined by the corral of atoms, but is also the waves seen traveling outward to and through the corral. Image produced by Crommie *et al.* (1993) and reproduced from Collins (1993) by permission of the American Institute of Physics.

a weight variation in the nucleus of the atom causes a small perturbation in the electron shell energies, leading to a subtle effect on the chemical and physical properties. Also, in the presence of gravity, natural processes tend to separate out isotopes of various types, an important effect in understanding aspects of Earth's history.

Isotopes of a given element may be stable, meaning that they have no tendency to change over time, or they may be unstable. An unstable isotope loses a portion of its nucleus (*radioactively decays*) through emission of particles of various types; very unstable isotopes may split apart. Some unstable isotopes last billions of years before they decay; others decay so rapidly that they are hard to study in the laboratory. The forces associated with the stability of the nucleus are discussed in Chapter 3; radioactive decay as a means of forming elements and dating cosmic events is discussed in Chapters 4 and 5, respectively.

In the discussion of chemical bonding and sharing of electrons, the reader may be left with a significant degree of dissatisfaction. How do electrons interact, and why do they preferentially move in certain patterns around atoms? To understand this requires that we free ourselves of the simple picture of elemental matter as particles. The behavior of microscopic atomic and subatomic particles displays attributes that have no real analogue in the macroscopic world.

An electron is a wave pattern, partly localized in space and energy. An electron around an atom will have a particular wave pattern or *wavefunction*, which is altered when another electron is introduced to complete the energy level. Through the extraordinary insights that led to quantum mechanics, such wave patterns can be calculated mathematically to understand how electrons and nuclei will interact to form atomic and molecular associations. However, intuition has yet to catch up: one simply must imagine that, at smaller and smaller scales, the discrete nature of matter finally loses its meaning in a sea of wave packets interfering one with another.

Technology today is allowing us to see the wave–particle duality of nature. Electrons can be manipulated to image individual atoms in a technique called *scanning tunneling microscopy*. In the image shown in Figure 2.9, IBM scientists arranged iron atoms in a circular pattern on a copper substrate and put a single electron in the center of this "corral". The single electron is seen not as a discrete particle trapped by the barrier of atoms but as a complex set of waves, extending beyond the corral.

There is much in the story of Earth that requires taking the microscopic, quantum view: in the next two chapters, for example, understanding the origin of the elements of which we are made, and the source of the light from the Sun, which has driven the evolution of our atmosphere and life. Our brief discussion of the microscopic world in the present chapter has not included many subatomic particles other than the ones described here. Indeed, one of these, the shadowy neutrino, has been invoked as one possible component of dark matter, and will be mentioned again in Chapter 4. To understand the evolution of the cosmos on the largest scales of space and time requires dealing with the quantum behavior of matter at its smallest scales.

Summary

The range of sizes in the cosmos is so enormous that scientific notation, a system of recording the number of powers of ten, is used to save space and aid multiplication and division of extremely large and extremely small numbers. The cosmos is not only large; it is also characterized by a range of motions from the spin of the Earth and the cycles associated with planetary orbits to the continued expansion of the cosmos as a whole. Agricultural societies throughout history have tracked the cycles of days, seasons, and years in order to ensure harvests and maximize productivity. But such efforts evidently went beyond agriculture into questions of the mechanistic nature of time and space. An ancient culture in England may have built the monument Stonehenge to track the occurrence of eclipses of the Sun by the Moon or of the Moon by the Earth, while the classical Mayan civilization of Central and South America saw in the cosmic cycles of planetary orbits a reflection of a cyclical nature to time itself. In modern times the scale of the cosmos has been assembled from a series of different techniques that apply to successively larger distances and which overlap so that one technique provides a bridge to the next. Thus the determination of the size of the orbit of the Earth allowed the angular shift in the positions of nearby stars as the Earth moves in its orbit, called parallax, to be translated into absolute distances to the stars. A convenient unit of measurement is the light-year, the distance light travels in vacuum in one year, which is about 10 trillion kilometers. The ultimate distance in the cosmos at which objects can be observed is limited by the speed of light, and corresponds to over ten billion light-years. At these scales the agglomerations of stars, called galaxies, themselves collected into groups called clusters and superclusters, are flying apart from each other. The most distant a galactic cluster or supercluster, the faster its velocity away from us. This discovery led to the concept that the universe began a finite time ago, in an explosion called the "Big Bang", and indeed the radiation left over from that explosion, moving away from us too, is measured primarily as a background radio static whose properties give an age to the cosmos of 13.7 billion years. Most of the cosmos seems to be made of unfamiliar matter, which does not reflect or emit light, called dark matter. The expansion of the cosmos – including the space between the galaxies – is accelerating under the influence of a very poorly understood property of reality called dark energy. On the microscopic level, normal matter is composed of elementary particles with different masses and electric charges. The chemical behavior of matter is governed by the number of negatively charged electrons associated with agglomerations of positively charged protons and uncharged neutrons. These assemblages are called atoms, and approximately 100 different kinds of atoms – called elements and distinguished by the number of protons – exist in nature and combine according to the nature of the interactions between electrons to form an enormous variety of different substances. A given element may have several different numbers of neutrons, which affect the mass of the atom; such different flavors of elements are called isotopes.

Questions

1. Construct a mental picture of the distances within the solar system by scaling the diameter of the Sun to the size of a soccer ball. What then would the distance from the Sun to Earth be? From the Sun to Jupiter? From the Sun to the nearest star?
2. What in the appearance of a crescent Moon, particularly in the evening or early morning sky, might be a clue to the fact that the Moon is spherical?
3. Our ability to measure parallax is limited by the size of the Earth's orbit. But spacecraft have been sent out to the edge of the solar system, to explore Pluto and the Kuiper Belt. How might such spacecraft be used to to enhance the measurement of parallax?
4. If the expansion of the universe is accelerating such that the most distant galaxies are receding over a horizon defined by the travel time of light from there to here, speculate on what the universe would look like to a far future intelligent species here in the Milky Way galaxy, say 100 billion years from now, when most clusters of galaxies have gone over this observable horizon thanks to cosmic acceleration. What kind of model of the universe might such beings adopt? How could they come up with the idea of a "Big Bang"?

5. The mental picture of electrons as charged balls whizzing around an atom is a very crude one, given that electrons have properties that are as much like packets of waves as they are like particles. Think of other examples of phenomena, even commonplace ones, that defy full description through common words.
6. The Large Hadron Collider (LHC) at CERN in Switzerland began operations in 2008. It collides beams of protons at speeds so high that the energies released in the collisions is 7 "Tera" electron volts. What is this energy in terms of common units of energy measurement, such as the energy needed to melt a kilogram of water? Using articles written about the LHC, describe what kinds of subatomic particles scientists hope to create there and what they hope to learn.
7. It is said that insight into the largest temporal and spatial scales of the cosmos will come from colliding the smallest particles together in powerful nuclear accelerators. What is meant by this statement?
8. Look at the periodic table. Which other element would you expect to most behave like carbon, from the point of view of chemical bonding? Why?

General reading

Arny, T. T. 2007. *Explorations: An Introduction to Astronomy*. McGraw-Hill, New York.

Davis, T. M. and Lineweaver, C. H. 2004. Expanding confusion: common misconceptions of cosmological horizons and the superluminal expansion of the Universe. *Publications of the Astronomical Society of Australia* **21**, 97–109.

de Pater, I. and Lissauer, J. J. 2001. *Planetary Sciences*. Cambridge University Press, Cambridge, UK.

Krupp, E. C. 2003. *Echoes of the Ancient Skies: The Astronomy of Lost Civilizations*. Dover Publications, Mineola, New York.

Randall, L. 2005. *Warped Passages: Unravelling the Mysteries of the Universe's Hidden Dimensions*. Ecco, New York.

References

Collins, G. P. 1993. STM rounds up electron waves at the QM corral. *Physics Today* **46**(11), 17–19.

Considine, D. M. (ed.) 1983. Chemical elements. In *Van Nostrand's Scientific Encyclopedia*. Van Nostrand Reinhold, New York, pp. 595–616.

Crommie, M. F., Lutz, C. P., and Eigler, D. M. 1993. Confinement of electrons to quantum corrals on a metal surface. *Science* **262**, 218–20.

Hartmann, W. K. 1983. *Moons and Planets*. Wadsworth, Belmont, CA.

Hoyle, F. 1972. *From Stonehenge to Modern Cosmology*. W. H. Freeman, San Francisco.

Linder, E. 2006. Seeing darkness: the new cosmology. *Journal of Physics Conference Series* **39**, 56–62.

Mason, S. F. 1991. *Chemical Evolution*. Clarendon Press, Oxford.

Robbins, R. R. and Jeffreys, W. H. 1988. *Discovering Astronomy*. John Wiley and Sons, New York.

Schele, L. and Miller, M. E. 1986. *The Blood of Kings: Dynasty and Ritual in Maya Art*. George Brazziller, Inc., New York.

Snow, T. P. 1991. *The Dynamic Universe: An Introduction to Astronomy*. West Publishing, St. Paul, MN.

Taylor, M. D. 1960. *First Principles of Chemistry*. D. Van Nostrand, Princeton, NJ.

Tegmark, M. 2007. Many lives in many worlds. *Nature* **448**, 23.

Tegmark, M. 1997. On the dimensionality of space-time. *Classical and Quantum Gravity* **14**, L69–L75.

3

Forces and energy

Introduction

The previous chapters have touched on the scale of the universe and the nature of the smallest pieces of matter. The structure of the universe is determined not just by the matter contained within it, but by the forces that both bind matter together and compel it to move apart. These forces, which act at the macroscopic and microscopic levels, are thought to be carried by certain types of subatomic particles. In the case of electromagnetism the force-bearing particle is called the *photon*.

We have learned most of what we know of the universe around us by studying the light coming from objects; our most information-filled sense is that of vision, and we have augmented it through the use of devices that can measure in detail the energy distribution of the light. This energy distribution from celestial bodies reveals much about their chemical composition and physical condition. Light from one such self-luminous body, the Sun, is the primary power source for Earth's climate and for life on the planet. The light by which the Sun and other stars shine is not generated by chemical reactions, but by reactions involving the nuclei of atoms at enormous pressures and temperatures deep within these gaseous objects' interiors; these are called *nuclear reactions*.

The nuclear reactions powering stars have, over time, generated essentially all of the natural elements except hydrogen, the most abundant element, and some of the helium (the remainder having been made from hydrogen in the primordial Big Bang). Thus the elements that make up life today (carbon, nitrogen, oxygen, phosphorus, etc.), with the exception of hydrogen, were manufactured by the very same process that today provides the energy source sustaining life on the planet. This chapter sets us on an evolutionary course that joins up eventually with the history of Earth and life, as we consider the processes by which elements are made.

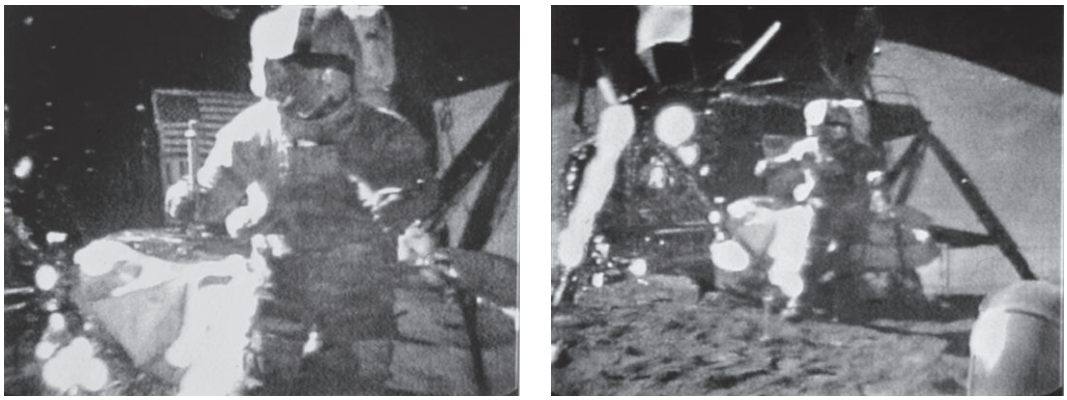
3.1 Forces of nature

Our lives are lived under the continual action of four forces that act in different ways upon matter. Two of these forces were deduced from the observation of everyday experiences; the other two act upon subatomic particles and were discovered and explored through laboratory experiments.

To discuss the nature of forces it is necessary first to define what a force is. This is surprisingly difficult, because we live constantly under the influence of forces (particularly gravity) that affect the paths of motions of objects. Thus, we are used to seeing a thrown baseball follow a parabolic trajectory under the influence of gravity, but in the absence of forces the ball would move with uniform velocity, that is, constant speed *and* direction, after it leaves the hand of the thrower. Thus, as first expressed by the seventeenth century English scientist, Sir Isaac Newton, in his extraordinary masterpiece *Principia*, every body continues in its state of rest, or of uniform motion in a straight line, unless it is compelled to change that state by forces impressed on it. An operational definition of *force*, then, is an action that causes a change in velocity (which could be a change in either speed or direction or both) of an object.

Closely related to force is acceleration, which is defined as the rate at which velocity changes. The reader may have inferred that velocity is a quantity that contains both speed and direction of a moving object. Thus, a car making a turn at a constant speed is accelerating, and its occupants feel a force on their bodies as surely as they do when the car is increasing or decreasing its speed in a constant direction. The force exerted by an object is acceleration multiplied by the object's mass.

The *gravitational force*, or gravity, is the attraction that all bodies exert on one another by virtue of their mass. The acceleration due to gravity is proportional to the mass of the attracting body. Because all objects have a gravitational force, one might say the attraction is mutual. However, for humans standing on Earth, the gravitational acceleration imparted to them



(a) (b)

Figure 3.1 Space-age version of Galileo's experiment on gravitational acceleration. (a) *Apollo 15* mission commander David R. Scott holds a hammer in his right hand and a falcon's feather in his left. (b) Having dropped both simultaneously in the airless environment of the moon – with no drag – both the heavy and the light objects hit the lunar surface at the same time, demonstrating that gravitational acceleration is independent of mass. Image from NASA television.

by Earth is much greater than the acceleration they impart to Earth.

A careful reading of the above definition reveals that a cannonball and a feather will be accelerated by Earth's gravity at the same rate. This seems counterintuitive, but our experience is "contaminated" by the effect of atmospheric drag on the less-massive feather. One can minimize the effects of atmospheric drag by using two balls of the same size but of different weights, and dropping them both at the same time, but a far more dramatic demonstration was conducted on the airless Moon in 1971. Apollo-15 commander David Scott, space-suited against the lunar vacuum, dropped a massive rock hammer and a falcon's feather brought from Earth simultaneously (Figure 3.1a); the television camera showed both reaching the ground at the same time (Figure 3.1b). Note that the gravitational *force*, however, is directly proportional to the mass of the object being accelerated. Thus, the hammer hit the lunar surface with much more force than the falcon's feather, in spite of the fact that they were being accelerated to the same extent by the lunar gravity.

Gravity is a so-called long-range force; it decreases according to the square of the distance between objects. On Earth we do not notice this, because the relevant distance is that to the center of the Earth. By moving up to the highest mountain (Everest), one moves only 0.15% of Earth's radius above its surface; thus the gravitational attraction of Earth decreases by only 0.3%. However, the force of Earth's gravity at the distance of the Moon, some 380,000 km (or 60 Earth radii) away, is 3,600 times weaker than at Earth's surface.

This inverse-square property of gravity is responsible for the characteristics of the orbits of the planets around the Sun (and of natural satellites, or moons, around the planets). Kepler's laws, which describe the elliptical shape and the property that the planet's path sweeps out "equal area in equal time" along the orbit, are both consequences of this property. Orbital motion is a balance between the gravitational force exerted by the Sun and the force associated with the changing velocity of the planets (and likewise for the Moon's motion about Earth and that of other natural satellites about their parent planets). Artificial satellites are launched into orbits around Earth by imparting to them a velocity sufficient to achieve a similar balance.

Tides arise from a particular effect associated with the distance dependence of the gravitational force and the fact that macroscopic bodies therefore will experience different forces at slightly different points in their interiors. The resulting tidal distortion can lead, under some circumstances, to stresses in the interiors of the planet and its satellite, which produce frictional heating of the interior. In the case of Earth and the Moon, tidal interaction causes the oceans of the Earth to slosh back and forth, which we see as the rising and lowering of the ocean along shorelines during high and low tides, leading to energy dissipation that slows the rotation of Earth and causes the Moon to gradually spiral outward to a larger orbit. (The ocean tides are modulated by the Sun as well, which is more massive than the Moon but much farther away.) The implications of the lengthening day are discussed in Part III.

The root cause of the gravitational force is poorly understood. In the context of *general relativity*, essentially a geometric theory of the origin and effect of gravity, the German physicist Albert Einstein visualized space as being distorted around objects, the extent of distortion being dependent upon the mass. Any physical object existing in space will have its path altered, or experience an acceleration, because of the distortion of space. Even the fundamental particles of light – photons – which have no mass, are predicted by this model to have their paths bent by gravity and this has been verified experimentally. Moreover, in relativity theory, time is a fourth dimension in the fabric of space-time: the theory predicts that the passage of time slows in the presence of a gravitational field, a prediction that also has been verified experimentally.

However, such a picture does not actually explain how matter interacts to produce the space–time distortions, and we must turn to a particle viewpoint: forces (including gravity) are assumed to arise by the mediation of special particles. A mass-bearing (or "massive") particle (neutron, proton, or electron, for example) emits a force-carrying particle. The resulting recoil changes

the velocity of the emitting massive particle, and the collision of the force particle with a second massive particle causes a velocity change in the latter. The properties of the force-bearing particle and the emitting and absorbing mass particles determine the strength and attractive or repulsive nature of the force. For gravity, the force-bearing elemental particle is called a graviton. This particle, a theoretical construct, has never been observed; however, other types of force-bearing particles have, leading physicists to hope that such a particle can be found for gravity.

The *electromagnetic* force is the force of repulsion or attraction that bodies with net electric charges exert on each other. Unlike the gravitational force, which is entirely attractive, electric charges come in two varieties – positive and negative – hence allowing two directions to the force: like charges repel; unlike charges attract. As with gravity, however, electromagnetism is a long-range force, decreasing as the square of the distance between bodies. Macroscopic objects, such as people, rocks, Earth, and the Sun, contain essentially equal numbers of positive and negative charges; hence we experience very little electromagnetic force. (Rub a balloon on a piece of fur, however, and a few charged dust particles accumulate on the balloon, allowing it to stick to walls.)

At the atomic level, where individual electrons are involved, the electromagnetic force dominates in chemical reactions (the sharing or exchanging of electrons) to form molecules. Furthermore, the physical properties of liquids and solids are dominated by the effects of the electromagnetic force associated with electronic attraction and repulsion. When we stand upon Earth, gravity pulls us to the center of the planet; we do not fall through the ground because the ground has solidity, and this in turn is due to the electromagnetic bonding of the atoms and molecules in the solid material.

The particle carrying the electromagnetic force is the photon. When a charged particle is accelerated, photons are emitted or absorbed. These photons carry electromagnetic energy through space, in a manner that is akin to waves traveling through a physical medium, such as sound waves through air. Electromagnetic waves, or trains of photons, can range in wavelength over arbitrarily large values. Those in the region of 5×10^{-5} cm, or 0.5 μm (micron), stimulate the human eye and are known as *visible light*. The electromagnetic force includes both electric and magnetic fields. A changing electric field induces a magnetic field, and vice versa. Earth and five other planets possess intrinsic magnetic fields that are generated by the motions of electrically conducting fluids in their interior (Figure 3.2). We detect the direction of Earth's field using magnetized iron, in the common device known as a compass. A few elements such as iron and nickel possess the property that they can be magnetized permanently by virtue of the tendency for certain kinds of alignments of their electrons. Such elements are *ferromagnetic*.

The *strong nuclear* force acts to attract protons and neutrons and hence to bind them into a nucleus, overcoming the repulsion between the like-charged protons. It is a short-range force the effect of which increases very sharply (*exponentially*, see Figure 3.3) as the distance between particles shrinks, but is negligible beyond about 10^{-13} cm. For this reason, nuclei tend to be less stable with increasing atomic number; some of the heaviest nuclei actually split apart or *fission*. The mediating force particle, called a gluon, is an exotic particle, evidence of which exists only in particle accelerator experiments.

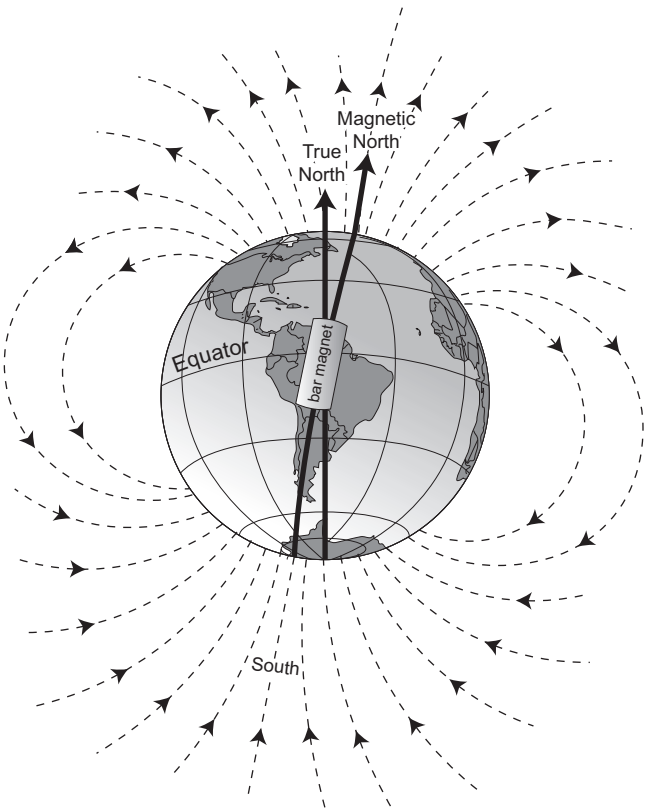


Figure 3.2 Visualization of the lines of magnetic force around Earth, generated by fluid motions deep inside our planet. Also shown is the misalignment between Earth's rotational axis and its magnetic axis. The shapes of the lines are valid close to Earth; farther away, the solar wind pushes the lines of force away from the Earth, creating a more complex magnetic structure. After Press and Siever (1978).

To understand the nature of the strong nuclear force, however, requires delving into the structure of the neutrons and protons themselves. They are not truly elementary particles, but in fact are composites of particles called quarks, which carry fractional electric charge. Three quarks are required to make up protons and neutrons, of two different types – "up" and "down." Protons are amalgams of two up and one down quark, whereas neutrons are two down and one up.

There are four other types of quarks, predicted by theory, which when compounded produce exotic massive particles normally found in particle accelerators (machines that collide subatomic particles at very high speeds) and extreme conditions in the cosmos; evidence for all six quarks has been found in accelerator experiments. The strong nuclear force, strictly speaking, binds quarks together, and in doing so creates a bound set of protons and neutrons that we call the *atomic nucleus*.

The trade-off between the influence of the strong force and the electromagnetic force provides a rationale for the number of protons and neutrons in naturally occurring stable isotopes. Sticking two or more protons together, in the absence of neutrons, is an inherently unstable exercise because the repulsive

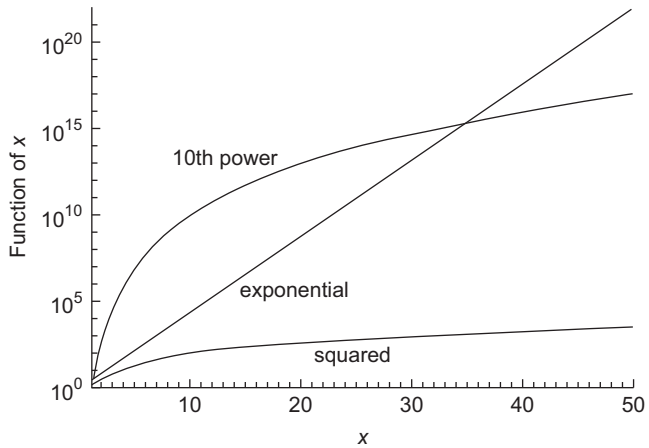


Figure 3.3 Functions that have an exponential dependence on some physical quantity increase (or decrease) much more quickly than those that depend only on some power of that physical quantity. Shown are three functions of a parameter x: x raised to the power 2, or x^2 (labeled "squared" on the figure); x raised to the power 10, or x^{10} (labeled "10th power"); and the "exponential" of x, or 2.7 raised to the power x. The exponential function is encountered again in Chapter 5, where radioactive dating is explained, and in Chapter 22 in the discussion of climate modeling. The exponential of x will always exceed x raised to any power for large enough values of x. On the vertical axis, values are plotted in scientific notation, and each tick mark represents a factor-of-10 increase from the tick mark below it.

electromagnetic force between the like-charged protons overcomes the attractive strong nuclear force. Inserting uncharged neutrons, which add to the attractive strong force, stabilizes the nucleus. As one moves upward in atomic number, larger nuclei formed of more protons and neutrons are less efficiently bound because the volume of the nucleus begins to exceed the effective range of the strong force. Hence a higher proportion of neutrons relative to protons is required to stabilize the nucleus (Figure 3.4) with increasing atomic number.

Eventually, beyond element 92 (uranium), the nucleus simply becomes so big that instability cannot be avoided. Heavier elements have been created by smashing nuclei together in nuclear reactors or particle accelerators. These *artificial elements* behave in exactly the same way as the naturally occurring elements; in particular, the electrons continue to systematically occupy higher energy levels (more distant from the nucleus) with increasing atomic number, as described in Chapter 2. The artificial elements tend to fission into lighter elements on short timescales. The distinction between artificial and natural reflects an Earth-centered bias, because some energetic processes elsewhere in the cosmos produce small quantities of the so-called artificial elements.

A prediction of the model of the nucleus is that some ultraheavy elements are stable. Somewhat analogous to electrons, neutrons and protons can be visualized as being organized within the nucleus in a series of concentric energy levels. As with the electrons, particular stability is achieved when levels are completely filled. Beyond uranium, the next stable region lies somewhere between 112 to 118 protons, and in 2007 scientists at the Joint Institute for Nuclear Research in Dubnya, Russia, were able to synthesize and study just two atoms of element 112 for several seconds before these broke apart in a process called radioactive decay, discussed in the next section. At the time of writing of this chapter, synthesis of element 114 has also been reported.

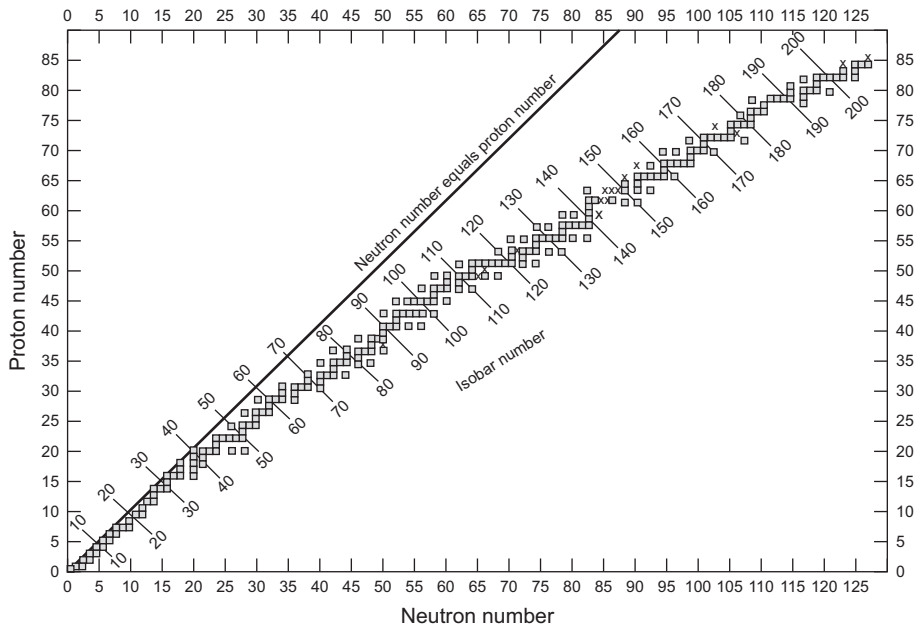


Figure 3.4 Distribution of stable atomic nuclei. The number of protons (atomic number) is plotted against the number of neutrons (atomic mass minus atomic number). For a given number of protons, and hence a given element, there are often several stable isotopes, and in some cases many. Beyond the lightest elements, stable nuclei tend to have more neutrons than protons. Some isotopes are labeled with an X; these are not strictly stable, but change very slowly over billions of years. The short diagonal lines define nuclei that have the same mass; that is, they have the same *total* number of neutrons and protons. This *isobar* number is important in the discussion of radioactive decay in Chapter 4. Redrawn from Broecker (1985).

The final of the four forces is the so-called *weak nuclear* force. It acts on electrons and more exotic atomic particles that are related to electrons. (These particles including the electron are not made of quarks, but are "elemental" on the same level that the quarks are.) The weak force also acts over a very short range, akin to the strong nuclear force. The weak force is manifest in an atomic nucleus when a neutron converts into a proton and an electron, with the electron leaving the atom. The result is that an unstable isotope of one element is converted into a different element with the same atomic weight as that of the decaying isotope. The mediating force particles are the so-called *massive vector bosons*.

3.2 Radioactivity

Radioactive decay refers to the spontaneous change of an atomic nucleus through emission of a particle, or splitting of the nucleus. Four types of radioactive decay can occur:

Alpha decay, or α decay, involves emission from the nucleus of two protons and two neutrons as an aggregate. Such an agglomeration, called an α particle for historical reasons, is in fact the nucleus of a helium atom, and is very stable. The original atom is left with a reduction of four in atomic mass, and two in atomic number, and hence is converted into a lighter element.

Beta decay, or β decay, involves conversion of a neutron in the nucleus into an electron and a proton. The proton stays behind, and the electron departs from the nucleus. This decay process, mediated by the weak nuclear force, leaves the atomic weight the same but advances the atomic number by one.

Gamma decay, or γ decay, does not alter either the atomic weight or the atomic number of the nucleus. A photon is emitted from a nucleus that has been put in an excited state (because of a collision or another decay process), a state in which the configuration of the protons and neutrons is at an elevated energy level. The loss of the photon decreases the energy of the nucleus, but the number of protons and neutrons remains unaltered.

Fission is the splitting of a massive atomic nucleus into two less massive pieces, forming two new elements of lower atomic number and weight than the original decaying element. Spontaneous fission involves release of energy as the nucleus splits. Fusion, the opposite process, involves the combining of lighter nuclei to form a heavier one.

Radioactive decay plays an important role over the history of the solar system in providing heat sources for planetary interiors, and natural chronometers in rocks through which the ages of important planetary and cosmic events can be dated. We discuss these further in Chapter 5. We return to fusion in Chapter 4 as the primary source of energy coming from stars, including the Sun.

3.3 Conservation of energy, and thermodynamics

A very important and universal concept of physics is *conservation of energy*. Simply put, energy is neither created nor destroyed, but only transferred from one form to another. Energy can be divided roughly into two forms: energy of motion, or *kinetic* energy, and energy stored in some fashion, called *potential energy*. Kinetic energy is straightforward to visualize; a moving car or falling stone both possess it. Kinetic energy is computed readily as half the mass of the object multiplied by the square of its velocity. Thus, when a car doubles its speed, it is quadrupling its kinetic energy; this is why the destructiveness of automobile accidents increases so dramatically with speed.

Potential energy is stored energy. The storage medium might be certain chemical compounds that, under the right conditions, tend to react in such a way as to release heat or exert pressure on a container; gasoline is an example. Storage of energy also can involve placement of material in a field that can induce movement; a weight suspended above the ground possesses potential energy, which becomes kinetic energy if the string is cut. A battery contains potential energy that can be released by creating a circuit of conducting wires connecting the two terminals.

Heat illustrates that energy can be considered as both kinetic and potential, depending on the context. Heat itself is in fact the random motions of molecules or atoms of a substance; each molecule can be thought of microscopically as having some kinetic energy by virtue of its motion. On the other hand, from the macroscopic point of view, a hot substance can be used to drive an engine, when there is a colder reservoir available to provide a direction of heat flow. In this situation the substance can be thought of as possessing potential energy by virtue of its heat content. Most important, though, is the concept that energy can be changed from one form to another and hence made to do useful work. Work itself usually is defined in terms of a change in the state of a system, for example, movement, increase in volume, or change in chemical composition. The rate at which energy is expended, for example in doing work, is defined as *power*. The basic metric unit of energy is the joule; a joule per second (power) is a watt.

Energy and matter are interchangeable; the sum of energy and the energy equivalent of matter is conserved. (This is a more general statement of energy conservation than the one given above.) Einstein's famous formula refers to the conversion of matter into an equivalent amount of energy equal to the original mass times the square of the speed of light in vacuum. A hydrogen bomb converts large amounts of mass into energy, by converting four hydrogen atoms into helium; the mass of each helium atom is somewhat less than that of the four hydrogens, and the "missing mass" goes into the energy of the explosion.

One fundamental concept related to heat is *temperature*. Temperature does not directly measure the heat content of a body, which also requires knowing the heat capacity, or how much heat can be stored in some object. For example, a very tenuous gas at high temperature may have less heat content than massive adobe walls at a much lower temperature. In fact, temperature relates to the average kinetic energy, and hence speed, of random collisions between the atoms of an object. Temperature is measured in a number of ways, the mercury thermometer being just one.

Temperature scales are important in this book. The Fahrenheit scale is commonly used but is awkward scientifically because 0°F doesn't correspond to *absolute zero*, or the point at which there is the minimum possible motion in a body (which is never

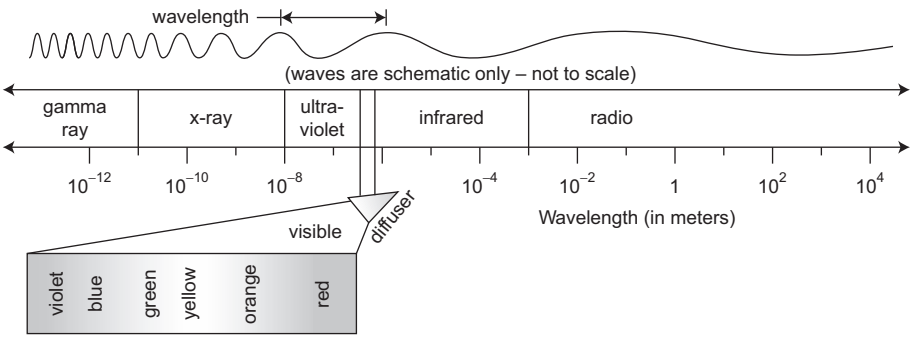


Figure 3.5 Schematic of electromagnetic spectrum, showing the names assigned to the various wavelength ranges. The visible part of the spectrum is expanded below to allow colors to be labeled. Wavelengths in meters are shown using scientific notation.

zero because of the intrinsic uncertainty in particle position and speed predicted by quantum mechanics and verified by experiment). The *Celsius* scale, used most places in the world, has the same problem. To get a Fahrenheit temperature, multiply the Celsius temperature by 9/5 and add 32. Thus, water boils at 100°C, or 212°F. A more rational scale, called the *Kelvin* scale, slides the Celsius scale so that 0 Kelvin is absolute zero. To do this, just add 273 to the Celsius temperature. Thus water, which boils at 100°C, does so at 373 K (the degree sign is not used in the Kelvin scale). Room temperature is roughly 300 K. The freezing point of water, at 32°F or 0°C, lies at 273 K.

3.4 Electromagnetic spectrum

What we perceive with our eyes as light is a particular form of electromagnetic radiation, as are radio waves, x-rays, and ultraviolet light. Electromagnetic energy propagates through space in the form of massless particles called photons, which we first introduced as the particles that mediate the electromagnetic force. Because photons are not fully localized in space, they also have the properties of waves of electromagnetic energy, which move through space as alternating electric and magnetic fields. Photons are created when free electric charges (such as electrons) are accelerated, or when bound electrons shift from one energy level in an atom to a lower one. Because of the wavelike properties of photons, electromagnetic radiation can be characterized by its *frequency*, the rate at which the waves pass a definite point, expressed as hertz or waves per second, and *wavelength*, the distance from one crest of the wave to another. The energy content of a photon is just proportional to its frequency.

Electromagnetic radiation travels through a vacuum at a definite velocity, 300,000 kilometers each second (3×10^5 km/s). In a material medium, electromagnetic radiation slows down, by an amount dependent on the frequency of the radiation. Devices called *spectrometers* or *spectrographs* therefore can be constructed to make the path traveled by light a function of wavelength, such that the various wavelengths are detected within different portions of the spectrometer, and the intensity (or number of photons) at each wavelength can be measured.

Figure 3.5 shows the names assigned to the different portions of the electromagnetic spectrum. It must be emphasized that these names are for convenience, are sometimes historical, but in no case imply a fundamental difference in the nature of photons from one end of the spectrum to the other. Photons at different wavelengths do interact with matter in very different ways, as is apparent throughout the book; nonetheless, the essential nature of photons as massless carriers of electromagnetic energy is the same throughout the spectrum.

Photons, in interacting with matter, can produce characteristic patterns or distributions of electromagnetic energy that reveal the physical and chemical nature of the matter. These *spectra* are most conveniently divided into three types: *continuum, emission,* and *absorption*. Continuum radiation is the broad distribution of photons characteristic of the temperature of a material, which in turn is just a measure of the mean speed of the atoms as they collide with each other. A material that is dense enough, or otherwise has the right properties, to absorb photons effectively and re-emit them will have a precise relationship between its temperature and the number of photons emitted at each wavelength. Such a material, called a *black body* if it is a perfect absorber of radiation, will exhibit a definite pattern of emission of radiation, called its *Planck function* after the German physicist Max Planck, who in 1918 received the Nobel Prize for his work on black-body radiation.

Figure 3.6 shows the pattern, or continuum spectrum, of radiation emitted by objects of different temperature. Light in the middle of the visible part of the spectrum is emitted by objects of temperature roughly 6,000 K, such as the Sun. Objects in a typically heated room, at 300 K, emit in the infrared. Microwaves are the peak of the Planck function for very cool objects (10 K), and gamma rays require objects in the million-degree range.

Note that only very hot objects create photons in the visible part of the spectrum, where we see with our eyes. Common objects around us, including ourselves, are visible because photons emitted by luminous sources such as the Sun or light bulbs are reflected from such objects and then are available to stimulate the retinas of our eyes. *Reflection* is simply the redirection of existing photons coming from a source. We see colors different from that of the source of illumination because physical materials selectively absorb photons at certain wavelengths and hence reflect only a subset of those incident upon them. The textures that we see in objects relate to how photons are reflected or scattered in different directions. Without the photons emitted from luminous sources such as the Sun, we would have no

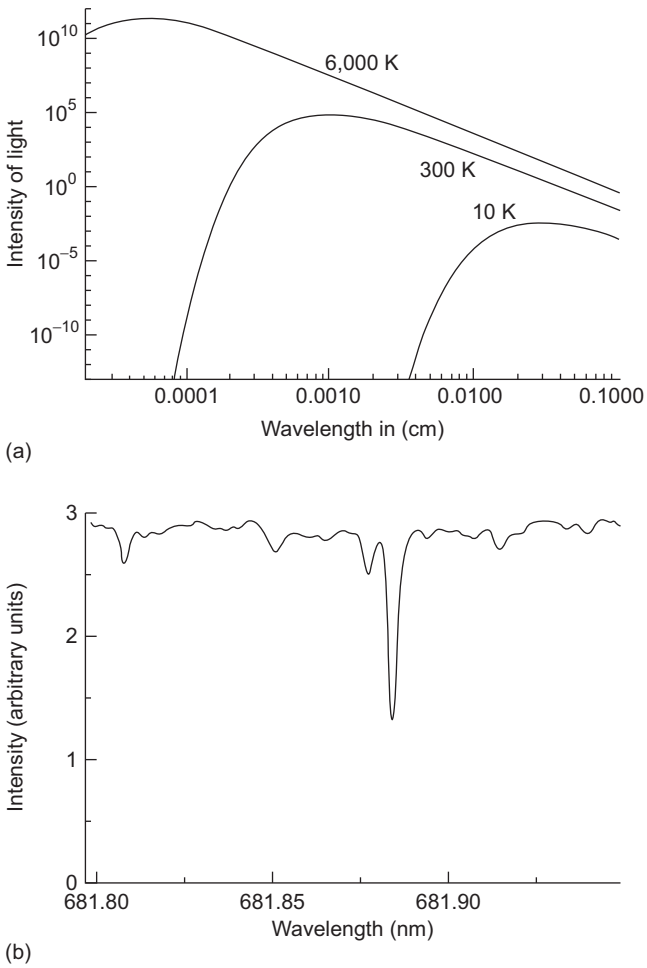


Figure 3.6 (a) Distribution of electromagnetic energy emitted versus wavelength for black bodies of temperatures 10 K, 300 K, and 6,000 K. The intensity of the photon radiation is expressed as energy emitted (in units of joules) every second at each wavelength, per square meter, per solid angle. The reader can appreciate, without worrying about the units, the fact that cooler objects radiate at longer wavelengths than warmer objects, and emit far less energy as well. (b) Example of a spectrum from the molecule methane, in wavelengths corresponding to red light, as measured in the laser spectroscopy laboratory of G. Atkinson at the University of Arizona. The wavelengths in this panel are given in nanometers (nm), or billionths of a meter (1 nm is 10 angstroms).

photons available by which to sense material objects around us. Moreover, as we discuss in later chapters, these photons provide the energy for warming Earth's surface and atmosphere, and the energy for most of the living systems existing on Earth.

Most materials are not perfect black bodies, and even those that are have this property only over limited wavelength ranges. Because of the electronic structure in atoms and molecules, with fairly well-defined energy levels between which electrons can move, electromagnetic energy tends to be absorbed or emitted at fairly definite wavelengths. If an electron drops from a higher energy level to a lower one, a photon is created, that is, emitted from the atom or molecule, at a wavelength or frequency characteristic of the difference between the two energy levels. A discrete, bright line will be seen at that wavelength if the light from the object is passed through a spectrograph.

The lines, or spectrum, emitted by an object that is undergoing electronic changes in energy level can be as intricate as fingerprints on a human hand, and just as diagnostic of the composition of that material (i.e., of what elements it is made); spectra taken in the visible and ultraviolet parts of the spectrum reveal such lines. Furthermore, in molecules, shifts of electrons from one major energy level into another are divided further into shifts between sublevels, associated with how the atoms in the molecule are vibrating relative to each other, and how the molecule is rotating. (Recall that something that rotates is accelerating, because it is not moving in a straight line; hence photons may be released or absorbed during rotation.) Spectra taken from the near infrared through to the microwave can reveal the identity of molecules through these complex transitions (Figure 3.6).

In very-high-temperature objects, such as the interiors of stars, collisions are large enough (and photons energetic enough) that atoms are partly or completely stripped of electrons. These *ions* then are positively charged particles, and the soup of positively charged ions and negatively charged electrons is called a *plasma*. Ionization occurs near the surfaces of hotter stars, and emission spectra from such ions are characteristic of how many electrons each atom has lost.

Complementary to emission spectra are absorption spectra, where photons are absorbed by atoms or molecules with associated increases in the energy levels of the electrons, or increases in the vibrational or rotational energy of the molecule. Such absorption spectra are, for our purposes, just the complement of emission spectra. Whether one finds absorption or emission spectra depends on whether a hot material is emitting photons into a colder region (*emission* spectra) or a colder object is absorbing photons from a warmer environment (*absorption* spectra) (Figure 3.7). In understanding Earth and the other planets, both kinds of spectra are important. What is key here is that the pattern of emission or absorption lines, even from a material composed of many kinds of molecules, can be deciphered to determine what molecules are present. Additionally, absorption or emission lines are more prominent when there are more atoms or molecules to absorb or emit photons at characteristic wavelengths; hence spectra provide information on the amount of atoms or molecules in a certain material.

3.5 Abundances in the Sun

One of the most striking first accomplishments of modern spectroscopy was the identification of a previously unknown element in the Sun. Beginning with the German scientist Josef Fraunhofer in the early nineteenth century, scientists had mapped dark absorption lines coming from the bright surface, or *photosphere*, of the Sun, and identified most of these lines with elements known at the time by measuring their spectra in the laboratory. A prominent line escaped identification until 1895 when the British chemist William Ramsey isolated a gas from a uranium-bearing mineral *cleveite*. Heating the gas in the laboratory produced the previously unidentified line in the solar

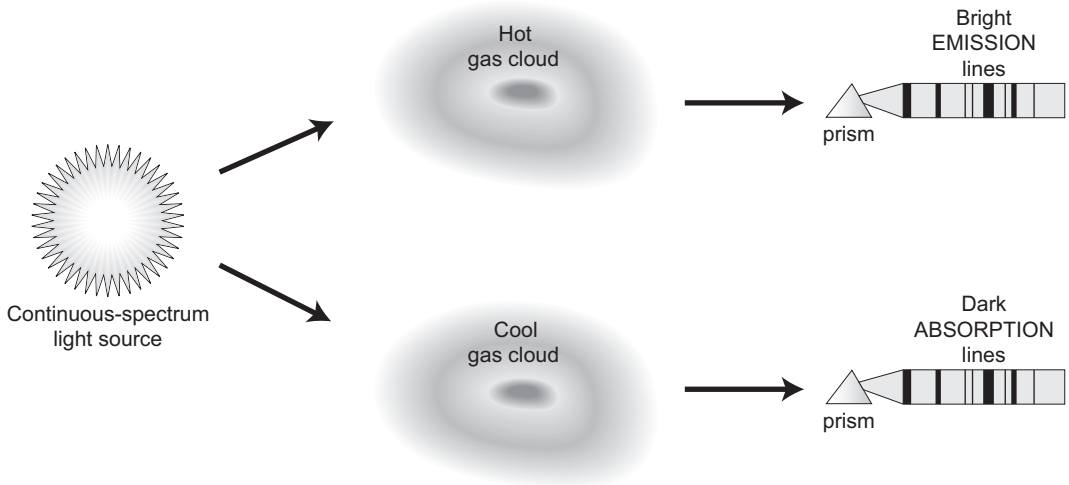


Figure 3.7 Emission spectra (top) are produced by hot substances emitting photons. Absorption spectra (bottom) occur when a flow of photons is intercepted, and absorbed at discrete wavelengths, by a cooler material. After Snow (1991).

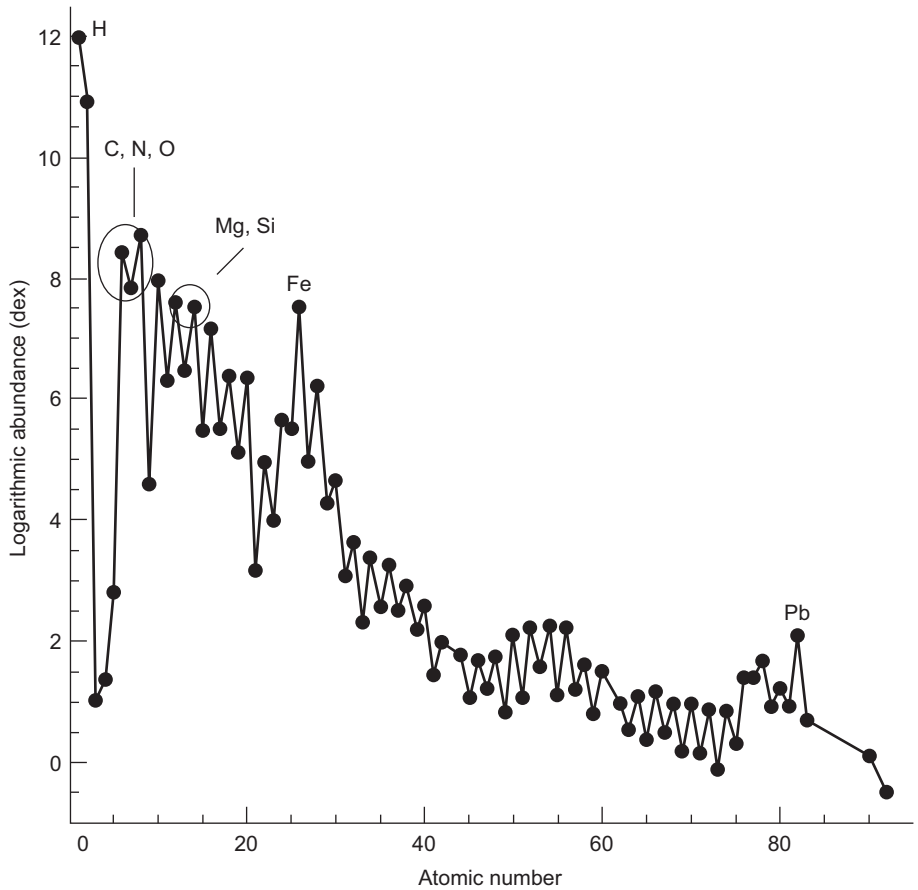


Figure 3.8 Primordial solar system abundances of the elements, derived from values in the Sun and primitive meteorites (see Chapter 5). Abundances are plotted on a *logarithmic* scale, so that each tick mark on the vertical axis means an increase or decrease by a factor of 10. Reproduced by permission from Broecker (1985).

spectrum, and the newly discovered element was named *helium* after the Greek word for the Sun (helios).

Since then, a thorough identification of abundances of elements in the Sun has been pursued, in part for its significance with regard to solar system objects. If the Sun and the planets had a common origin, as we argue in Chapter 11, then, because the Sun has retained essentially all of its gas from that time, it should contain the original, primordial mix of material from which the planets formed. There are difficulties in deriving a complete inventory from spectroscopy, however. One only sees the surface of the Sun, and to assume that the interior abundances are equal to those on the surface requires the notion that the Sun is thoroughly mixed. This is not entirely true, and corrections must be made. In addition, the nuclear reactions that power the Sun (Chapter 4) convert some elements to others, and corrections must be made for this also in deriving an inventory of "original" planetary material.

Abundances of elements in the Sun are summarized in Figure 3.8. Most are derived from spectra of the Sun's surface, but additional information is folded in. This includes direct sampling by spacecraft of the *solar wind* (a stream of charged atoms emanating from the Sun) and material from dust emitted by comets, and chemical analysis of a class of meteorites (rocks that originally were in orbit around the Sun and collided with the Earth) that are thought to be relatively unaltered since the birth of the solar system (Chapter 5) and hence have some elements in their original relative abundances.

The resulting graph is a guide to answering some important questions about how the planets, including Earth, have evolved over time, and such problems are discussed in later chapters. Perhaps more profound, however, is that the pattern of elemental abundances reveals something about *how* these elements were formed in the first place. The decline in abundance toward higher proton number, the peak near iron, and then a further decline, and the zigzag nature of the abundances for odd and even proton numbers reflect the superposition of a number of natural nuclear reactions that have taken place since the birth of the universe as a whole, and to which we turn in Chapter 4.

Summary

Matter acts under the influence of what appears to be four forces in nature, which are called the electromagnetic, strong nuclear, weak nuclear, and gravitational forces. At least the first three of these forces appear to have a common origin and at very high energy should behave as a single force. In contrast, the force of gravity is much weaker than the other three and in most theoretical models of the nature of forces has so far defied unification with the other three. Each force can be thought of as being carried by a "virtual" particle that exists for a brief time over which a force acts on matter. In the case of the electromagnetic force, this particle is the familiar photon, which is the unit or quantum of light – or more generally, electromagnetic energy. The electromagnetic force is repulsive between two particles of like charge and attractive between particles of opposite charge. Magnetism is another manifestation of the electromagnetic force. Gravity is an attractive force between any two masses, which declines as the distance squared between the two objects. The force of gravity was the first force to be investigated in controlled experiments: Galileo Galilei's demonstration that two objects of different mass will experience the same acceleration under the force of gravity. The strong nuclear force binds together protons and neutrons, making the atomic nucleus stable, but acts only over very short distances. A nucleus that is too large or contains too many neutrons, will therefore split in some fashion in a process called radioactive decay. The weak force acts on electrons and exotic subatomic particles related to the electron; it too is manifested in a form of radioactive decay in which a neutron spontaneously converts into a proton and an electron. In addition to the properties of matter and the behavior of the forces of nature, the physical universe is governed by conservation laws, the best known of which is the conservation of energy. Energy can be converted from one form to another, or converted to matter and back to energy, but is never truly created out of nothing or destroyed. The predictability of the cosmos under the properties of matter, energy, and forces is best illustrated by the nature of the electromagnetic spectrum. Light, as does matter, has properties that are both wavelike and particle-like. Thought of as a wave, the wavelength of light determines how it interacts with matter; the longest wavelengths of light are radio waves while the shortest are gamma rays. The shortest wavelengths have the highest energies – hence gamma rays can destroy the structure of a material like biological tissue whereas radio waves will merely pass through without causing damage. Electrons in atoms generally interact most strongly with light at optical wavelengths – where our eyes can see – and shorter, ultraviolet, wavelengths. Assemblages of atoms called molecules will be caused to vibrate or rotate by absorbing light at infrared wavelengths. The nature of the interaction of light with electrons bound to atoms can be thought of as quantized – discrete amounts of energy are absorbed or emitted by atoms and molecules, causing a characteristic increase or decrease in the intensity of light emitted or reflected by an object at particular wavelengths. Thus the spectrum or

distribution of light with wavelength coming from an object – whether it be self-luminous, reflecting, or absorbing the light from another source – is characteristic of the composition of that object. The composition of the Sun, and later other stars, could be deduced by comparing their spectra with spectra of elements and molecules studied in the laboratory. Thus the abundances of the elements in the cosmos have been determined through the observation of light from stars, and the pattern of abundances then provided astronomers key clues to how such elements were formed – deep within the interiors of the stars themselves.

Questions

1. What are the various conversions from one form of energy to another that take place as an automobile engine is started and then engaged to propel the automobile along a road?
2. What is the original source of energy for the gasoline that is used to power the automobile? (Hint: check Chapter 23.)
3. Why does the particle nature of light seem to be most manifest in the gamma ray part of the spectrum, whereas light at radio wavelengths behaves more like a wave?
4. Knowing that the charges of the quarks must add up to zero for the neutron and plus one for the proton, the reader might have fun deducing the fractional charges on the up and down quarks.
5. The spectra of stars from the hottest to the coolest is classified according to the letter sequence O, B, A, F, G, K, M. What is the origin of this sequence? (You will need to consult text or web articles on astronomy.)
6. A simple molecule like molecular hydrogen – two hydrogen atoms – has a limited number of ways it can rotate and bend in response to its interaction with light. A more complicated molecule like methane – a carbon with four hydrogen atoms – can bend and twist in many more ways. Based on these considerations, what would you expect the differences would be in the general appearance of the infrared spectra of the two molecules?

References

Arons, A. B. 1990. *A Guide to Introductory Physics Teaching*. John Wiley and Sons, New York.

Boorstein, D. J. 1983. *The Discoverers*. Vintage Books, New York.

Brumfiel, G. 2007. Unseen universe: a constant problem. *Nature* **448**, 245–8.

Eichler, R., Aksenov, N. V., Belozerov, A. V. *et al*. 2007; Chemical characterization of element 112. *Nature* **447**(72), 47–9.

Hogan, J. 2007. Welcome to the dark side. *Nature* **448**, 240–5.

Mason, S. F. 1991. *Chemical Evolution*. Clarendon Press, Oxford.

Pippard, A. B. 1957. *The Elements of Classical Thermodynamics*. Cambridge University Press, Cambridge, UK.

Poppy, W. J. and Wilson, L. L. 1965. *Exploring the Physical Sciences*. Prentice-Hall, Englewood Cliffs, NJ.

Press, F. and Siever, R. 1978. *Earth*. W. H. Freeman, San Francisco.

Snow, T. P. 1991. *The Dynamic Universe: An Introduction to Astronomy*. West Publishing, St. Paul, MN.

4

Fusion, fission, sunlight, and element formation

Introduction

The understanding of the origin of sunlight (and starlight in general) was a nineteenth and early twentieth century development that culminated in the release of nuclear energy in human-made devices on Earth. Beyond the implications (both negative and positive) of such developments, however, lies the profound perspective gained in the latter half of the twentieth century regarding the origin of the elements of the periodic table. The existence and abundances of the 90-odd elements that make up Earth, the planets, the solar system, and the universe beyond have an explanation that lies in natural nuclear reactions that have taken place in the several generations of stars preceding the formation of the Sun and the solar system.

4.1 Stars and nuclear fusion

The observable cosmos around us is, by and large, made of stars. Stars are spheres made primarily of hydrogen and helium gas; the size of the spheres is determined by a balance between the attractive force of gravity pulling everything inward and the pressure associated with the high temperatures of stars' interiors, which is a force tending to push the material outward. Most stars eventually evolve, through nuclear processes described below, into dense spheres of carbon, oxygen, or exotic neutrons; some collapse into the mysterious and incredibly dense *black holes*.

The copious amounts of photons coming out of stars, including the Sun, are a signature of the enormous temperatures in their interiors. The origin of these high temperatures, and hence of sunlight or starlight, was a matter of debate throughout the nineteenth century. A hypothesis by the British physicist Lord Kelvin, that the Sun was radiating away the energy associated with its initial collapse from clouds of interstellar gas and dust, met with a timescale problem: the Sun would cool in several tens of millions of years, but various lines of evidence suggested that terrestrial rocks were older by at least a factor of 10. However, the essential and simple concept that the infall of material by gravity toward a common center, forming a star or planet, would generate heat is essential to understanding the heat budget of Earth, as we discuss in Chapter 11. Another possible source, radioactivity of heavy elements, was advanced around the same time, but the spectroscopic determination that the Sun is mostly nonradioactive hydrogen and helium made this hypothesis also untenable.

By the 1930s, physicists began to grasp the essential workings of the atomic nucleus, including the fact that with sufficient force, one could overcome the repulsive barriers between the nuclei of atoms and induce lighter nuclei to combine to form heavy nuclei, in a process called fusion. In the case of four hydrogen nuclei (each of which is just a single proton) combining together, the most stable resulting nucleus requires that two of the protons transform to neutrons. This is accomplished only through a modestly complex series of steps, outlined below, but the important point is that the resulting nucleus *has less mass than the original four protons*. The missing mass ΔM has been converted to energy ΔE, according to the Einstein formula $\Delta E = \Delta M c^2$.

Analogous to electrons, certain numbers of protons and neutrons assembled as nuclei represent especially stable structures. In general, the stability of the nonradioactive nuclei increases as the atomic number increases toward iron; beyond iron, the stability tends to decrease. Therefore, fusion reactions tend to produce energy as heavier nuclei are assembled, only up to iron (Figure 4.1). Nonetheless, this does not mean that it is easy to fuse two nuclei together; sufficient pressure (or collisional force, and hence temperature) is required to overcome first electronic repulsion and then repulsion associated with the two colliding nuclei.

Production of heavier elements from lighter ones by fusion in stars appears to be a process of fundamental importance to the evolution of the cosmos and in particular to the existence of solid planets. It is therefore worth getting a flavor for the kinds

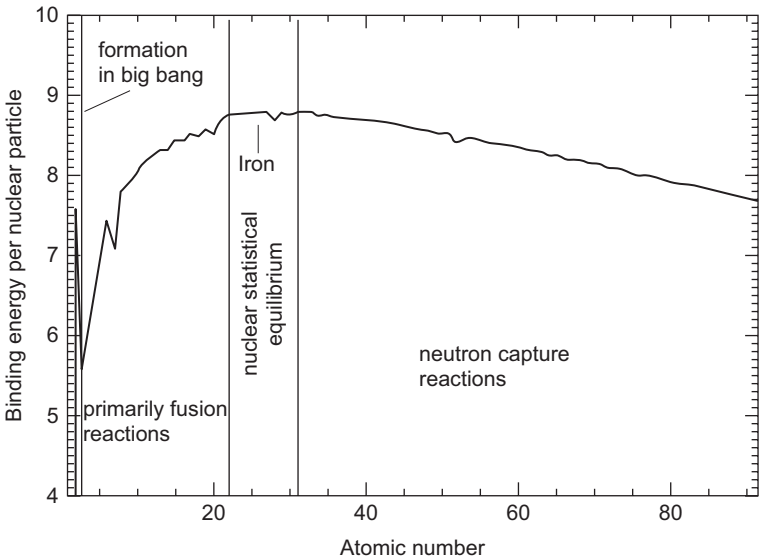


Figure 4.1 Binding energy of the nucleus as a function of atomic number. The higher the energy, the more stable is the nucleus against fragmentation or other decay. Note that stability is highest around the atomic number corresponding to iron. The binding energy is expressed relative to the number of protons and neutrons in the nucleus, and the units are millions of *electron volts*. One electron volt is 1.6×10^{-19} joules, and is a convenient unit for energies on the small scale of atoms. The curve was computed from a model that roughly fits the measured value for most elements, but there are small deviations from the experimental values. Vertical lines delineate the regions within which elements are produced (a) primordially, from the Big Bang; (b) via thermonuclear fusion inside stars; (c) through reactions at very high temperature during the stellar collapse that engenders a supernova explosion; and (d) neutron capture. Each of these is explained in the text.

of reactions that take place. We focus on the fusion of hydrogen to heavier elements. The simplest and most basic fusion process is called the *p–p chain*, or proton–proton chain, and requires that only hydrogen and helium be present.

The simplest of the p–p chains, often called ppI, involves three separate *reactions*, as sketched in Figure 4.2. A reaction is defined as a discrete step in the process in which one or more atomic nuclei are fused to form certain products. In ppI, step 1 involves two hydrogen nuclei (protons) colliding to form a deuterium nucleus (one proton and one neutron) and two atomic fragments. One such fragment is identical to an electron in mass, but of opposite charge, and is called a *positron*. Also released is an exotic particle with little or no mass and a propensity for passing easily through matter. Such *neutrinos* have been detected experimentally.

In step 2, the deuterium nucleus collides with another hydrogen nucleus, i.e., proton. The net result is the release of light, or photons, and the generation of the two nuclei into a light isotope of helium, consisting of two protons and one neutron. This isotope, helium-3 (^{3}He), is quite rare in the cosmos, because it is destroyed easily by further fusion reactions. However, some of it escapes from the Sun in the solar wind, and has been detected.

Finally, two ^{3}He nuclei collide and form the most stable isotope of helium, helium-4 (^{4}He), consisting of two protons and two neutrons. This nucleus stays intact under present conditions in the Sun's interior but undergoes further fusion in more massive stars. In addition to the ^{4}He, two protons (i.e., the hydrogen nuclei) are produced.

If we take all of the reacting nuclei and the product nuclei from the three stages of the ppI chain, we see that, in net form, four protons have been converted into one helium nucleus, which consists of two protons and two neutrons. A net loss of mass has occurred, and that lost mass (a small fraction of the total) is liberated as energy, mostly as photons, and generates the light we see coming from the stars.

The ppI chain is not the only proton–proton fusion reaction chain that occurs in stars. Indeed, it is possible for ^{3}He to collide with a ^{4}He nucleus to form beryllium, and from there lithium (ppII chain) and, in a fraction of reactions, boron (ppIII chain), but in the end these heavier elements are destroyed in favor of ^{4}He again. Beryllium and lithium act as *catalysts* in the nuclear reaction; they control the speed of the reaction sequence, in this case by being good targets for electrons and protons, without being consumed in the process. We see examples of catalysts in *chemical* reactions (that is, involving whole atoms as opposed to just the bare nuclei) in Chapter 13 but the principle is the same as with the nuclear reactions discussed here.

The energy liberated per helium nucleus produced is 4×10^{-12} joules, enough to power a 40-watt (40-joule per second) light bulb for only 10^{-13} seconds. However, the Sun contains enough hydrogen to produce 10^{56} helium atoms; if all of the hydrogen were so converted, the amount of energy released would be 4×10^{44} joules. The Sun is observed to emit photon energy, its main form of removal of the energy, at a rate of 4×10^{26} watts; therefore, the p–p chain could sustain this process for 10^{17} seconds, or 30 billion years. This calculation is a bit off because (a) much of the hydrogen is too far from the Sun's center to experience high enough temperatures to undergo fusion and (b) the Sun's brightness (luminosity, or power) has varied with time. More careful calculations yield roughly 12 billion years of steady hydrogen fusion for the Sun, of which 4.5 billion years has already transpired.

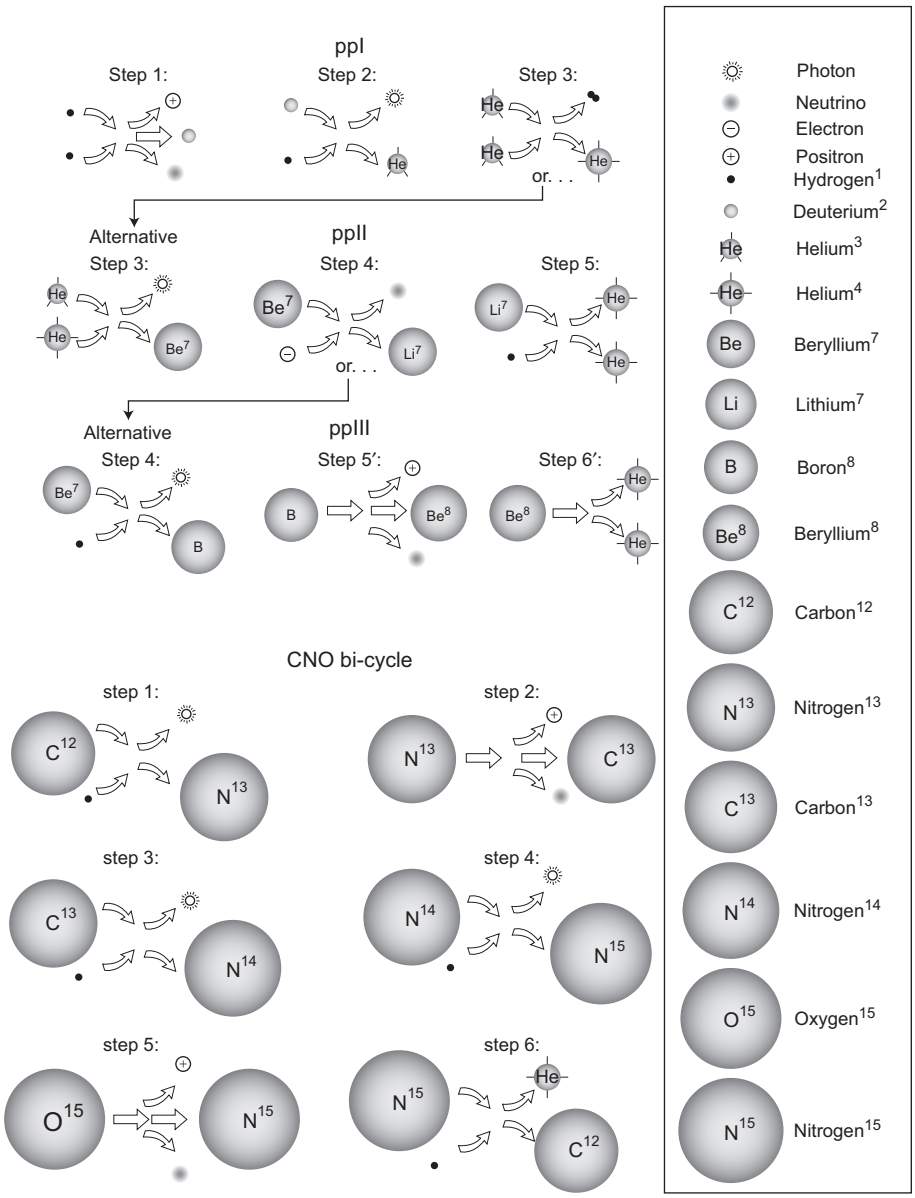


Figure 4.2 Steps involved in four kinds of fusion reactions in stars, all of which convert hydrogen to helium.

The p–p chain is only one of two cycles converting hydrogen to helium in stars. The carbon–nitrogen–oxygen, or *CNO* cycle, requires that the three heavier elements, so familiar to us on Earth, be present in the region of nuclear burning. In this cycle, carbon acts as a catalyst to facilitate, through the intermediate formation and destruction of nitrogen and oxygen, the creation of the ^{4}He nucleus from four protons. The sequence is potentially much faster than the p–p chain because, in the latter chain, two hydrogen nuclei (protons) must collide to initiate the process, and this is inefficient because of the small size of the protons. In the CNO cycle, all collisions are between protons and larger nuclei such as the carbon nucleus (six protons and six neutrons). However, there is so little carbon in the center of the Sun that the CNO cycle is currently less important than the p–p cycle. As fusion proceeds in the Sun and helium builds up, the interior temperature of the Sun will increase; as the temperature increases, the CNO cycle will gain in importance and eventually dominate.

Fusion requires very high temperatures to provide nuclei with enough velocity to overcome the repulsive electric force between the protons. In stars, the high temperatures are achieved through the enormous pressure associated with the mass of the star: our Sun is 1,000 times more massive than Jupiter and 300,000 times more massive than Earth. Most stars, however, are smaller than the Sun, and in the interiors of the smallest stars, or *red dwarfs*, nuclear reactions barely proceed, and are much slower than in the Sun. These stars are cooler, and hence appear red rather than yellow, but they are much longer lived because they burn hydrogen more slowly. Stars more massive than the Sun undergo hydrogen fusion much more rapidly, are much brighter and bluer than the Sun, but are far shorter lived. During the time over which stars undergo hydrogen fusion, their brightness and

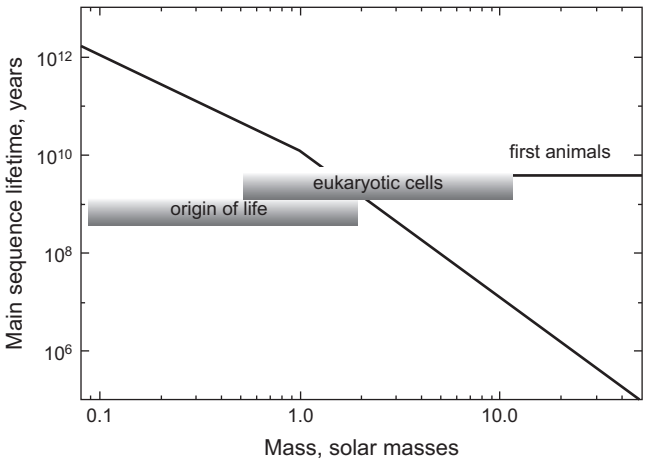


Figure 4.3 Main sequence lifetime of stars over which they stably undergo hydrogen fusion, as a function of mass of the star expressed in solar masses. The Sun's mass on this scale is 1. The lifetime is given in years. The timescales for the appearance on our planet of life, of complex cells, and of animals are shown. Note that the lifetime and mass must be expressed in powers of 10 because of the broad range of stellar masses and ages.

size change only slowly; this stable portion of their evolution is referred to as the stellar *main sequence*.

The lifetime and luminosity of main sequence stars, sorted according to their mass, have important implications for the habitability of orbiting planets and the chance that life will have enough time to evolve into complex forms before these stars become unstable. Figure 4.3 shows the time for stable hydrogen fusion in stars as a function of their mass; this is the main sequence lifetime. Stars several times more massive than our Sun do not last long enough to give complex life a foothold on any planets orbiting them, if the timing of evolution of life on our planet is a fair guide (Chapter 12).

Like normal hydrogen, deuterium is being depleted today by fusion processes in stars. in fact, deuterium can undergo fusion at lower temperatures than can hydrogen. The reaction is simple: it is the second step of the ppI chain in which a deuterium nucleus and hydrogen nucleus collide to form a ^{3}helium nucleus with liberation of energy in the form of a photon. The reaction of two protons, the first step of the ppI chain, requires much higher collision velocities and hence limits hydrogen fusion to objects more massive than those that just undergo deuterium burning. The threshold for hydrogen fusion in stars of solar composition is 85 times the mass of Jupiter; for deuterium burning it is only 13 times Jupiter's mass. Because of the very small abundance of deuterium in the cosmos – 50 parts per million relative to hydrogen – deuterium fusion can only power a star for a few million years at most, compared to the billions of years that stars such as the Sun shine by hydrogen fusion.

4.2 Element production in the Big Bang

Hydrogen fusion produces helium, which builds up as a kind of thermonuclear "ash" in the interiors of main sequence stars. Stellar explosions, which we discuss below, can deliver helium to the gas in interstellar space, the *interstellar medium*. This production of helium from hydrogen is just one example of *stellar nucleosynthesis*, or the production of elements within stars. It is a somewhat special case, however, because, unlike most of the elements, much of the helium present today in the cosmos is thought to be primordial, like hydrogen. The origin of the primordial material is presumed to lie in the initial explosion that started the universe, that is, the Big Bang.

Evidence for an initial explosion of matter to create the cosmos exists primarily in the observed expansion of groups of galaxies away from each other, and in the pervasive background static, mentioned in Chapter 2, which can be heard at radio wavelengths. This background static is produced over a range of wavelengths, and the energy is distributed so as to be a nearly perfect black body with a temperature of 2.7 K. The most straightforward interpretation of this cosmic static is that it is the last light from the initial explosion, red-shifted by its great distance from us, marking a transition from a universe that in its first moments was suffused with photon radiation scattering off of a dense gas of subatomic matter.

During the initial phases of the expansion after the Big Bang, the universe consisted mostly of neutrons compressed to an extremely high density, much like the present-day interiors of *neutron stars*, the remnants of stellar collapse. Very quickly, however, β decay created a population of electrons and hydrogen nuclei, that is, protons. Helium was formed in this dense soup of matter through capture of a neutron by hydrogen to form its heavy isotope, deuterium, followed by collision between two deuteriums to make tritium (the next heavy, and unstable, isotope of hydrogen) plus hydrogen, and terminating with a collision between tritium and deuterium to make ^{4}He. A branch-off in the step involving two colliding deuterium nuclei produces ^{3}He, some of which survives today as a primordial remnant.

The Big Bang production of helium was a different process than the p–p chain in stars, emphasizing that different ambient conditions (in this case, the much higher densities in the Big Bang than are obtained in stars) force the nuclear reaction sequence to be different. In addition to most of the present-day helium coming from Big Bang nucleosynthesis, it is thought that most lithium available today was made at that time.

4.3 Element production during nuclear fusion in stars

Fusion reactions beyond hydrogen burning in stars require increasingly higher temperatures because the nuclei, up through iron, are progressively more stable (Figure 4.1). Consider the Sun as a typical star of intermediate mass and age. The temperature near the Sun's center is computed from physical models to be roughly 1.5×10^7 K, fully adequate for hydrogen fusion, but a temperature of 10^8 K is required to initiate the next stage in which helium fusion takes place. However, as more hydrogen is consumed, leaving a core of helium, the density increases and the temperature rises. Computer models suggest that temperatures near the center of the Sun are already 10% higher than they were at the time hydrogen fusion was initiated. The continued slow increase in the Sun's internal temperature leads to

increasing luminosity, which will have profound consequences for Earth's habitability well before the end of our star's main sequence lifetime.

As the Sun approaches the end of its main sequence life some 6 billion years hence, hydrogen fusion will become progressively concentrated in a shell around the then-helium core. This core no longer will be supported by the heat from fusion reactions, and will begin to contract rapidly, heating up to the threshold temperature for helium fusion. Two helium nuclei, that is, α particles, each composed of two neutrons and two protons, will collide to form ^{8}Be. This is an unstable isotope of beryllium, with a large cross section, and hence will easily capture another helium nucleus, producing the most abundant carbon isotope, carbon-12 (^{12}C).

Although the energy production of helium fusion is small compared with that of hydrogen fusion, the sudden pulse of ignition will force the Sun to expand dramatically and become a *red giant*, which will extend out beyond the orbit of Venus, almost to Earth. The luminosity of the Sun will increase sufficiently to bake the Earth and melt the water ice of Jupiter's Galilean moons.

Helium burning produces heavier elements by capture of additional α particles, each succeeding element having a mass number 4 higher than the previous element. Thus, α-particle capture by carbon produces ^{16}O; capture by ^{16}O produces ^{20}Ne, and ^{24}Mg then is produced from neon. In addition to this rather simple production by helium nuclei, which explains the high abundance of elements with mass numbers divisible by 4, other reactions are going on. The CNO cycle during hydrogen fusion has produced other elements such as ^{14}N. Addition of α particles leads to heavy isotopes of oxygen (^{18}O), neon (^{22}Ne), and magnesium (^{25}Mg). Other products of the CNO cycle are converted to ^{13}C, ^{15}N, and ^{17}O.

Further cycles of nucleosynthesis occur as helium is exhausted and the star again goes through core collapse and resulting heating. At a temperature of about 1 billion Kelvin, carbon fusion is initiated. Two carbons combine, releasing a helium nucleus and producing ^{20}Ne or releasing a proton and yielding heavy neon (^{23}Ne). Isotopes of magnesium, sodium, aluminum, and silicon are produced in the carbon fusion process as well. Oxygen nuclei also undergo fusion to produce isotopes of silicon, phosphorus, sulfur, magnesium, and aluminum.

Finally, temperatures exceeding 4 billion K permit silicon fusion to take place. These enormous temperatures occur only in the collapse of the most massive stars, and allow a variety of nuclear reactions to occur rapidly, creating iron and elements of similar atomic mass. This is the end of element production by fusion in stars, however, because beyond iron nuclei decrease in stability; fusion thus requires energy input and is not self-sustaining. Other processes associated with fusion, described in the next section, produce elements not directly made by fusion.

Only the more massive stars make it all the way to silicon burning; the Sun and smaller stars will only progress through helium fusion, after which final collapse will produce a small *white dwarf star*. For stars massive enough to produce iron by silicon fusion (nine or more times more massive than the Sun), the end is more dramatic: the termination of silicon fusion and core collapse compress a star's interior intensely, producing a violent *supernova* explosion. It is within the extreme environment of such an explosion, as well as other exotic astrophysical environments associated with massive stars, that additional element formation occurs.

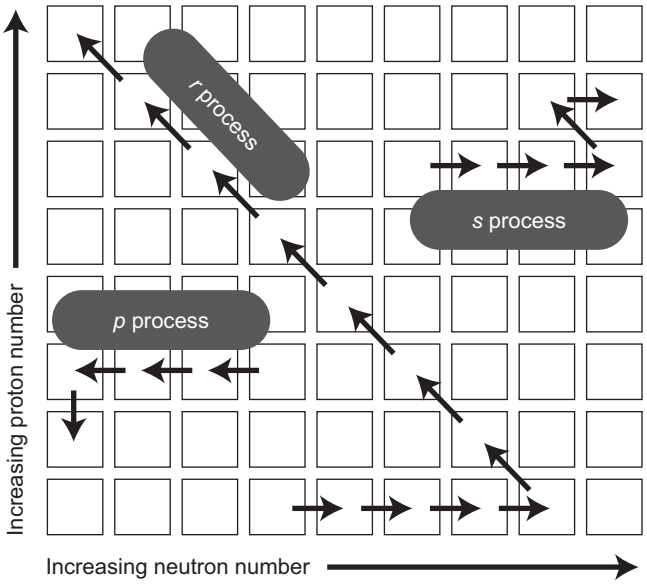


Figure 4.4 Graph of s, r, and p processes. Elements and isotopes exist in squares defined by a proton number (vertical axis) and a neutron number (horizontal axis). Straight horizontal arrows to the right are neutron captures; diagonal arrows represent β decays. Because of its complex nature, the p process cannot be shown directly but is equivalent to the horizontal line moving to the left accompanied by a downward vertical step. Adapted from Broecker (1985).

4.4 Production of other elements in stars: *s*, *r*, and *p* processes

The deep interiors of stars undergoing fusion reactions are dense fluids of protons, neutrons, electrons, and heavier nuclei (composites of protons and neutrons). In addition to the direct fusion reactions considered above, the capture of protons and neutrons by nuclei can build up the atomic mass (and in the case of proton capture, the atomic number) in ways distinct from the main fusion reaction sequences. Neutron capture is much more likely than proton capture, because electrostatic repulsion does not have to be overcome. Sources of free neutrons become important in the helium burning stage of a star's life. The production of ^{25}Mg from ^{22}Ne and an α particle, and the conversion of ^{13}C and an α particle to ^{16}O both liberate neutrons and are thought to be their primary sources.

The process of neutron capture in a stable stellar interior is the s, or slow, process. It is so defined because the flux of neutrons is such that the time between successive captures of neutrons by a nucleus may range from 10 to 10^5 years. An understanding of neutron capture is greatly aided by the sort of diagram shown in Figure 4.4. The graph shows the various elements and their isotopes, collectively called the *nuclides*, plotted as the number of neutrons on the horizontal axis versus the number of protons, defined earlier as the atomic number, on the vertical axis. Thus, a horizontal movement on the graph is from one isotope to another

of the same element; a vertical movement is from one element to another. The atomic mass of a species is given by the sum of the neutron number and the proton number. *Isobars*, or species of the same atomic weight, lie on a diagonal line from lower right to upper left on the chart.

Capture of a free neutron moves an isotope horizontally along the graph, converting it to a heavier isotope of the same element. Eventually, the isotope reaches a neutron number that is not stable, and decays radioactively. Because of the long time between neutron captures in the *s* process, such an unstable nucleus will undergo radioactive decay before the next capture, and the relevant radioactive process is β decay, in which a neutron converts to a proton and an electron, but the atomic mass of the nucleus (number of protons and neutrons) is preserved. On the graph, such an event moves the isotope diagonally up and to the left, along the isobaric line. β decay continues until a stable nucleus is reached, and then continued neutron capture moves the isotope, now a different element, horizontally to the right again.

The resulting abundances of elements and isotopes are determined both by the neutron flux and the relative cross sections of the various nuclei created. As mentioned earlier, the stability of nuclei depends separately on the numbers of both neutrons and protons. Certain of these numbers, as with electrons, are particularly stable whereas others are not. This is in addition to the unstable situation of having too many neutrons relative to the number of protons, leading to β decay. Very stable nuclei have small cross sections for capturing neutrons and hence tend not to be converted to heavier isotopes or isobars. The limited rate of neutron addition relative to β decay forces the pattern of diagonal movement along an isobar as soon as an unstable isotope is reached. Thus, although the *s* process is important in making many elements and isotopes above iron, it cannot produce the more neutron-rich isotopes.

The question of which stellar environments are the most important contributors of *s*-process elements is a continuing debate. Presumably, the *s* process goes on in all stars undergoing fusion beyond the hydrogen stage, but we are interested in stars from which material eventually is expelled in sufficient quantities that it is an important contributor to the interstellar medium and, eventually, to new generations of stars and planets. *Asymptotic Giant Branch* (AGB) stars swell in the late stages of nuclear burning and consist of a core of carbon and oxygen that is not undergoing fusion, surrounded by a shell undergoing helium fusion and a final, outer hydrogen layer. These stars appear to be abundant and might be important sites for *s*-element production.

Not all heavy elements can be made by the *s* process. Some neutron-rich isotopes require that neutron capture proceed quite far to the right, through the unstable isotopes, before β decay takes over. Rapid addition of neutrons, or an *r process*, is required. Here, capture of neutrons is rapid enough that very neutron-rich nuclei are produced, until the binding of additional neutrons becomes so unfavorable that the net capture rate is no longer competitive with β decay, and a cascade of β decays moves the neutron-heavy elements diagonally to the left in Figure 4.4 until a stable nuclide is reached.

Once one understands the stability of the various nuclides, charting their production by the *s* and *r* processes becomes a kind of board game in which the pieces are moved according to rules determined by nuclide stability, neutron fluxes, and the ambient physical conditions, elucidated through laboratory experiments and computer models. But what environment could be so neutron-rich as to enable the *r* process to occur? Stars several times more massive than the Sun that have completed fusion cycles up through production of the iron-group elements explode as supernovas. Neutron-rich environments within the rapidly expanding envelopes of supernovas have been invoked as possible sites for production of elements by the *r* process, but none seems capable of producing the full mix of *r*-process elements seen in the galaxy. The problem is an intricate one because not only must conditions be right for *r*-process element production, but the material then must be ejected into interstellar space without being further altered significantly.

A exotic, neutron-rich wind coming from a "neutron star" might be an additional site of the *r* process. After the explosion of a star as a supernova, the remnant cinder collapses with no further prospect of fusion reactions to halt the collapse. If the star is massive enough, collapse will continue "forever" and a black hole will be formed. Most supernova remnants, however, stop collapsing when the pressures are high enough that all electrons and protons are squeezed together to make neutrons. This incredibly dense neutron star is only a few kilometers across, yet it contains potentially as much mass as the Sun. For the first 10 seconds or so of its existence, an intense wind of neutrons flows from the neutron star, and it is in this cosmic neutron breeze that many or most of the *r*-process nuclides might be produced. The rate of neutron star births is thought to be high enough to make these winds a primary source. The reader should regard this model not as the last word, but as an illustration that the search for the birth sites of the elements is tied closely to an understanding of the exotic processes by which stars evolve and die.

Some nuclides in Figure 4.4 are relatively proton rich and are shielded from *s*- and *r*-process production by other stable nuclides. Some 35 nuclides out of the hundreds of stable and near-stable nuclides known to exist are in this state. For some time it was thought that a *p process* to produce such material must involve addition of protons. This is difficult because high temperatures are required to produce sufficiently energetic collisions for protons to overcome the electrostatic repulsion of other protons. Appropriate environments for proton addition within stars were difficult to find.

An alternative mechanism that enriches protons in a nucleus is removal of neutrons. To make the *p*-process nuclides, the removal would have to occur from stable nuclides, ones for which β decay will not operate. Exposing nuclides to very high temperatures for short periods of time is one possibility, because the neutrons will "drip" off of the nuclides first, followed by protons; if the process is truncated early enough, the net result is relatively proton-rich nuclides. Certain regions of the interiors of supernovas have been identified as providing the right environment for the *p* process, in which the supernova shock itself provides a short high-temperature burst. At least two different kinds of supernovas appear to be required to produce the right mix of the *p*-process nuclides, and it is clear that much more work will be required to fully understand how these are formed.

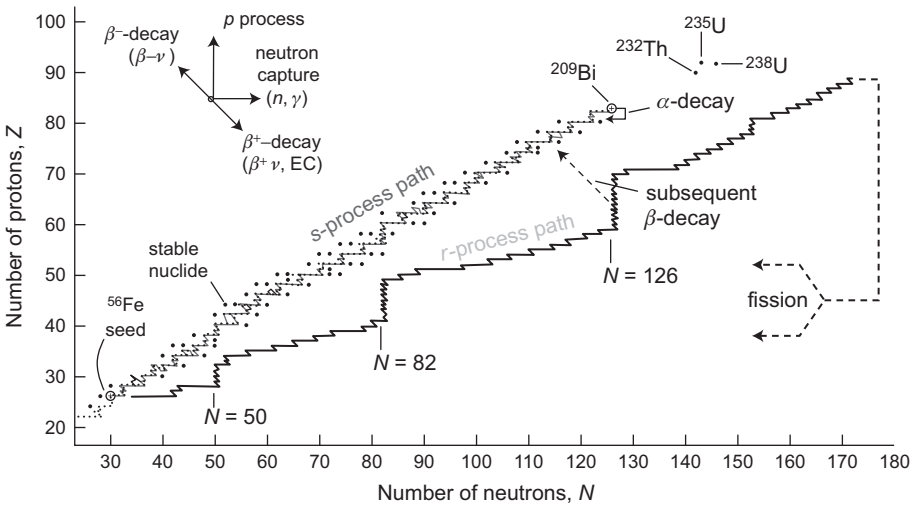


Figure 4.5 The processes of neutron addition plotted on a graph of proton (atomic) number versus neutron number. Beginning with iron and nearby elements, the more regular *s* process follows a zig zag line as neutrons are added and beta decay occurs, the path corresponding to "valleys" of high stability of nuclei with given proton and neutron numbers. The *r* process, on the other hand, adds neutrons so rapidly that the products are neutron rich, truncated by larger cascades of beta decay until the heaviest nuclei simply split through atomic fission. The *p* process is sketched as well for comparison with the neutron addition processes.

4.5 Nonstellar element production

Once expelled into interstellar space by supernova explosions or the more quiescent ejection of envelopes around lower-mass stars, element production and evolution are not terminated. Most nuclei that are ejected from supernovas have initial velocities an appreciable fraction of the speed of light. Nuclei that intersect our solar system and hit Earth are called high-energy *cosmic rays*. Collisions between ambient interstellar hydrogen and the high-energy nuclei cause spallation or splintering of portions of the heavy nuclei. This *l process* is a primary one in the production of lithium, beryllium, and boron.

Additionally, once produced, isotopes not fully stable begin to radioactively decay, which is another kind of element and isotope production process. Decay times range over large values, from seconds through billions of years. As described in Chapter 5, the abundance of decay products of some of these isotopes, trapped in rocks, provides a wealth of information ranging from the age of the solar system to the timing of geologic events on Earth.

4.6 Element production and life

Figure 4.5 provides a larger scale view in atomic number and neutron number space of the neutron addition processes that operate in astrophysical environments. Beginning with the elements around, and including, iron as seeds, neutrons are added either slowly or rapidly until beta decay converts neutrons to protons. Ultimately the production of the heaviest elements is truncated by fission of the nucleus or alpha decay in the case of the *r* and *s* processes, respectively.

The extent to which it is possible to understand the sources of elements and their isotopes is remarkable, given that only a century ago scientists were still struggling with the concept of the nature of elements and the underlying structure of the atom. Today we have a glimpse of the wide range of processes – from the Big Bang through stellar fusion and supernova explosions – responsible for the mix of elements present today in the cosmos.

It is particularly intriguing to examine the elemental abundances and notice that the fundamental building blocks of life – carbon, hydrogen, nitrogen, and oxygen – are quite abundant relative to most other elements. Except for hydrogen, which is the primordial element, these others are abundant because they are direct products of the fusion reactions powering stars.

The high abundances of silicon and iron-group elements have planetary implications. Silicon is the last of the source materials for main fusion reactions, the products being iron and elements close to it. These elements of moderate atomic weight are the basic building blocks, with oxygen, of Earth and its sister terrestrial planets; the compounds of such elements are loosely referred to as rocks and metals.

Go out into the dark skies of a moonless night in the countryside and gaze at the multitude of stars. Let your eyes run from the seven sisters of the Pleiades to the red giant Betelgeuse in the constellation Orion. In this visual sweep, one captures the alpha and omega of element production: young stars just beginning their conversion of hydrogen to helium by fusion, and the red giant going through its terminal stages of fusion before the frenetic final neutron production of heavy elements. There in the sky are the cosmic factories making the elements that, in the distant future, might become part of some strange biology on an as yet unformed world.

Summary

Normal stars are spheres of mostly hydrogen and helium, held together by the force of gravity, and heated by their original collapse to very high temperatures in their interiors. The high temperatures translate to vigorous random motions and high-speed collisions deep in their interiors. The collisions create an outward pressure balancing the inward force of gravity, and also strip electrons off of the atoms to create bare atomic nuclei or ions. Stars more than about 80 times the mass of Jupiter, or 25,000 times the mass of the Earth, have interiors at temperatures so high that the collisions are sufficiently energetic to cause nuclear fusion – a process whereby hydrogen is concerted to helium with release of energy. The process of fusion is actually a sequence of nuclear reactions involving the splitting off and recombination of various atomic particles. One set of nuclear pathways from hydrogen to helium, called the p–p chain, occurs predominantly in stars the size of the Sun and smaller, while the so-called CNO cycle occurs in more massive stars. A star's structure can be stably sustained by hydrogen fusion for millions of years in the more massive stars to trillions of years in the smallest, "red dwarf", stars. As hydrogen is converted to helium, the star's interior becomes denser, the temperature goes up, and the reaction rates increase. Eventually stable hydrogen fusion is no longer possible and the star expands, then contracts in several cycles, leaving the stable "main-sequence" of hydrogen fusion and undergoing additional cycles of fusion of heavier elements to produce carbon, oxygen, and heavier elements. Fusion ceases to generate energy for element numbers at and above iron, and so the most massive stars will cease fusion as iron is produced, collapsing catastrophically and blowing off much of their mass in the form of a supernova. The remnant core may be a dense clump of exotic neutrons or a black hole. Stars the mass of the Sun never reach this stage, ending as white dwarfs rich in carbon and oxygen, which cool slowly over cosmic time. The stages of stellar evolution after the main sequence may also be responsible for the production of elements not directly produced by fusion, or which are heavier than iron. In this way, most elements are produced during the life cycle of stars. The formation of the cosmos in the Big Bang produced hydrogen, some helium, and lithium, so that the first generation of stars were bereft of the heavy elements needed to make planets and organic molecules for life. It is thus the progressive formation of heavy elements in the interiors of stars, their expulsion into the cosmos at the end of the stellar main sequence, and recycling through later generations of stars, that has produced the mix of elements we see today in the cosmos.

Questions

1. Given the story of element production described in this chapter, would you expect life to have been possible during the very first generation of stars after the Big Bang?
2. Why might one not expect to encounter intelligent life on a planet orbiting a star twice the mass of the Sun?
3. It is said that if the relative strengths of the fundamental forces were slightly different than they actually are, fusion and element production would not be possible. Do a literature search to find the details behind this statement.
4. Speculate on the final demise of stellar nucleosynthesis in the far future: based on how much hydrogen has been converted to heavier elements since the Big Bang, how long might it take for hydrogen to become too rare a commodity for stable fusion to occur. Could "helium stars" be generated by collapse of helium-rich interstellar gas? What would the minimum stellar mass be (roughly) for such helium-burning stars?
5. Which hydrogen fusion process would not have been possible in the earliest history of stellar evolution, and why?
6. Explain why, for the lighter elements, the abundances are higher for those with an even number of protons.
7. Red dwarf stars undergo fusion at a slower rate, and hence are less luminous than the Sun typically by a factor of 100 to 1,000. If a planet orbiting a red dwarf is to receive as much starlight per second as the Earth receives from the Sun, how much closer to its star must the planet be than the Earth is to the Sun (pick either factor given in the previous sentence)?

References

Aldridge, B. G. 1990. The natural logarithm. *Quantum* **1**(2), 26–9.

Broecker, W. 1985. *How to Build a Habitable Planet*. Eldigio Press, New York.

Cloud, P. 1988. *Oasis in Space: Earth History from the Beginning*. W. W. Norton, New York.

Clayton, D. D. 1968. *Principles of Stellar Evolution and Nucleosynthesis*. McGraw-Hill, New York.

Mason, S. F. 1991. *Chemical Evolution*. Clarendon Press, Oxford.

Meyer, B. 1994. The *r*-, *s*- and *p*-processes in nucleosynthesis. *Annual Review of Astronomy and Astrophysics* **32**, 153–90.

Reiforth, R. 2006. Stardust and the secrets of in heavy-element production. *Los Alamos Science* **30**, 70–7.

Sackman, J., Sackman, I-J., Bootnroyo, A. I., and Kraemer, K. E. 1993. Our Sun III. Present and future. *Astrophysical Journal* **418**, 457–68.

Truran, J. W. Jr. and Heger, A. 2004. Origin of the elements. In *Treatise on Geochemistry V. 1*, ed. A. M. Davis. Elsevier Pergamon, Amsterdam, pp. 1–15.

Wilford, J. N. 1992. Scientists report profound insight on how time began. *New York Times*, April 24 **CXLI**(48, 946), p. 1.

Wilson, T. L. and Reid, R. T. 1994. Abundances in the interstellar medium. *Annual Review of Astronomy and Astrophysics* **32**, 191–226.

PART II

The measurable planet: tools to discern the history of Earth and the planets

5

Determination of cosmic and terrestrial ages

Introduction

To understand the history of Earth in the cosmos, we must be able to establish ages of physical evidence and timescales over which processes have occurred. The task is daunting because of the enormous spans of time over which the physical universe and Earth have existed, and several different approaches must be used. In Chapter 2, we discussed observations leading to the conclusion that the universe is in an overall state of expansion, which began some 13.7 billion years ago. In this chapter we discuss rather precise techniques that enable us to determine the age of the Earth and other solid matter in the solar system with even higher accuracy and perhaps more confidence: some 4.5682 billion years ago, the planet we live on began to take shape in the form of tiny solids condensed from a hot, gaseous disk.

5.1 Overview of age dating

It is useful to distinguish between two kinds of chronologies that are constructed in regard to Earth's history, because the techniques and uncertainties are quite different. A *relative chronology* is derived by observing the order in which a series of objects is found – and then assuming that the series represents a temporal ordering. In sediments on Earth, older layers of soil, sand, and rock are by definition those which are deposited first, hence they lie at the bottom of a sequence of layers progressing upward from oldest to youngest. If there is no disturbance, one can reasonably assume that the layers have been preserved in the order in which they were deposited. Geologic processes might turn a whole stack of layers upside down, but fossils present in the layers, which can be compared to those in other layers worldwide, enable us to determine the age progression of the layers and hence their inversion by some geological event. We discuss relative geologic dating in Chapter 8.

Similar relative records of events can be read from the surfaces of planets; on the Moon we find evidence, discussed in Chapter 7, of an early period of frequent impacts on the surface to form craters, followed by extensive volcanic flooding to make the lunar mare. On Mars, dried-up river channels are seen to be overlain by impact craters in some places, but cut through preexisting craters in others. Such photographic evidence allows a relative chronology of events to be constructed. In cases in which the average rate of physical processes can be estimated, relative ages can be assigned rough absolute values; however, this is an approach fraught with potential error.

Absolute chronologies, our main concern here, contain information on the actual times at which events took place. To construct such chronologies requires a natural and well-calibrated clock, with markers indicating when the "ticking" began. On a macroscopic level, biological growth effects such as rings in trees or seasonal events such as thawing of lake water provide clocks of greater or lesser accuracy; we encounter these much later. Certain microscopic processes, atomic or nuclear, have the simplicity and predictability required to act as very precise clocks over enormous time spans. Radioactive decay of certain isotopes of the elements (defined in Chapter 3) provides both the regularity and the markers required for such measurements, and as we see below, there is a broad range of radioactive nuclides characterized by varying longevity that occur in natural materials. Scientists have applied these to problems ranging from the age of ancient settlements (^{14}C dating) to the time when element formation first began in our galaxy (using long-lived uranium and thorium isotopes, among others).

5.2 The concept of half-life

To understand how radioactive isotopes, introduced in Chapter 3, can be used to date the materials within which they are found, we must delve into a little physics and mathematics. We have talked in Chapter 2 about quantum mechanics and the consequent *probabilistic* nature of atomic processes.

Table 5.1 Half-lives of important radioactive elements

Parent	Daughter	Half-life (Year)
^{235}U	^{207}Pb	0.70 billion
^{238}U	^{206}Pb	4.5 billion
^{232}Th	^{208}Pb	14.0 billion
^{40}K	^{40}Ar	12.0 billion
^{40}K	^{40}Ca	1.4 billion
^{87}Rb	^{87}Sr	49.0 billion
^{147}Sm	^{143}Nd	106.0 billion
^{26}Al	^{26}Mg	700,000
^{14}C	^{14}N	5,730

Imagine a single atom that is radioactive. Although one might know, from its particular identity as a radioactive isotope of a given element, whether it is likely to decay sooner rather than later, it is not possible to predict how long before it will decay, even approximately. This would seem to contradict the notion that radioactivity provides a precise clock for dating events.

However, because decay is a probabilistic event, precision is achieved through considering an ensemble of a large number of atoms of the same isotopic species at once. This is easy to do, because macroscopic materials contain enormous numbers of atoms. The rate of decay of a large number of atoms of a given radioactive species can be measured quite precisely. One way to express this rate is in terms of the *half-life*, which is simply the time it will take half of a sample of radioactive atoms – the "parent" – to decay to a stable "daughter" species when the number of atoms is very large. While not all parents decay directly into a stable daughter product – ^{238}U decay involves 14 intermediate steps and daughter products before reaching ^{206}Pb – the half-life is a measurable and dependable characteristic of a particular radioactive isotope system (Table 5.1) – not influenced by changes in pressure, temperature, or the composition of the environment around it.

Another important aspect of the radioactive decay process is that the number of decays in a given time is just proportional to the number of radioactive atoms present. This makes sense because, for a decay to take place, there must be radioactive atoms present, and the more present, the more decays that are likely to take place in a given amount of time. In fact, over a very short time interval δt, where δ indicates a discrete change in a quantity, a simple algebraic equation describes the change δN in the number N of radioactive atoms of a particular element present:

$$\delta N = -NR\delta t.$$

Here we use R to represent the rate at which the radioactive decay occurs. R is the reciprocal of the mean lifetime of the radioactive atoms, which is 1.4 times the half-life. Note also that the three quantities on the right-hand side – the number of atoms, the rate, and the time interval – are to be multiplied together. As is common in scientific writing, we do not put multiplication signs ($\times$) between the symbols. The minus sign is needed to indicate that the decay process decreases the number of atoms over time.

The equation only tells us the change in isotope number over time. If we want to know what the number N is as time passes, for example, over 10 days, we could add the increments δN over the time increments δt that total 10 days. Because N changes at each time step, we must make our steps sufficiently small that we accurately track N. There is a mathematical procedure, called *integration*, that we can perform to find N:

$$N(t) = N(0)e^{-Rt}.$$

In this equation we introduce two new types of symbols. $N(t)$ and $N(0)$ are simply shorthand for the number of radioactive atoms at times that we label t and 0, respectively. Our time 0 is arbitrary; depending on the situation that we are calculating, $t = 0$ might be last Wednesday at 9 A.M. or it might be the moment that Earth began. But, it makes sense that, to compute the number of radioactive atoms at time t, we must know what that number is at some earlier time.

More peculiar is the symbol e. It is shorthand for exponential, and it is a special function that produces a unique number for every value of $-Rt$. The value of e^1, or just e, is 2.71828 . . . , with the ellipses indicating that the number is not exact as written, but continues to run on indefinitely. Then, e^2 is e multiplied by e, or 7.38904 . . . . For negative numbers, we just take reciprocals: e^{-2} is $1/e^2$, or 0.135335 . . . . Things get a bit more complex when we have fractional powers for the exponent, that is, $e^{-3.45}$ for example, but this represents a number also, which can be worked out on a scientific calculator or computer. An exponential curve, e^x, which rises very steeply as x increases, is displayed in Figure 3.3.

Why do we end up with this exponential function describing radioactive decay? It is because the number of atoms decaying is proportional to the number of radioactive atoms present. If this were not the case – if, for example, the number decayed δN per time interval δt were just proportional to R – then the decay law would be simple: $N(t) = N(0) - Rt$. However, many physical processes, of which radioactive decay is but one – growth of bacteria (because each bacterium present splits into two), initial growth of a fertilized egg, etc. – operate in such a way that the change in a quantity depends on how much of that thing is available. Such processes are referred to as *exponential* in their growth, or *inverse exponential* if there is a negative sign in the power, as in radioactive decay.

Figure 5.1 shows the progressive, inverse exponential decline of radioactive atoms during a decay process. It also shows when the half-life is reached – after half of the initial atoms have decayed. The radioactive decay law is fundamental to what follows in this chapter, though it is by no means the whole story, as we shall see. Nonetheless, the predictability of the decay of an ensemble of a large number of atoms is at the crux of the use of this process as a clock.

In the remainder of the chapter we consider two different approaches to dating materials by radioactivity, distinguished by what actually can be measured in the system. The first of these is *radiocarbon dating*, limited, because of the short half-life of radioactive ^{14}C, to organic remains of living things that died less than about 70,000 years ago. The second technique is applied to radioactive isotopes with much longer half-lives, such that both the amount of the original radioactive isotope, hereinafter the *parent*, and the product, hereinafter the *daughter*, species can

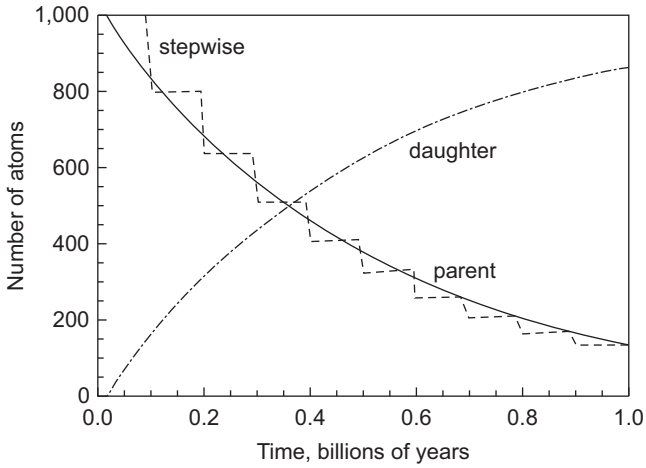


Figure 5.1 Radioactive decay law. A sample of 1,000 atoms (a very small number compared to real samples analyzed) is assumed, which have, for example, a half-life of 355 million years. The numbers of parent atoms remaining, and daughter atoms produced, are shown as a function of time. The half-life can be read from the graph as the time corresponding to the crossing of the parent and daughter curves. The dashed line is the number of parent atoms remaining based on simply adding the increments δn; this stairstep pattern only roughly approximates the real decay law.

be measured or inferred. This technique is used in the dating of terrestrial and extraterrestrial rocks many millions or billions of years old.

5.3 Carbon-14 dating

The stable isotope carbon-12 (^{12}C) is one of the more abundant atoms in the cosmos, and a foundation for biology on Earth. Carbon-13 (^{13}C) also is present as a stable isotope in all natural carbon-bearing systems, but at much lower abundance. The next heavier isotope, carbon-14 (^{14}C), is continually produced in Earth's atmosphere as the most abundant nitrogen isotope, ^{14}N, absorbs neutrons produced from an influx of atomic fragments – the *cosmic rays* from energetic sources in the galaxy. The absorption of the neutron leads to ejection of another neutron or a proton, but primarily the latter. When the proton is ejected, the atomic number decreases by one but the mass stays at 14, and hence ^{14}N is transformed into ^{14}C (Figure 5.2).

As Table 5.1 indicates, ^{14}C decays with a half-life of approximately 5,730 years. The production of ^{14}C by neutron bombardment and its decay lead to a roughly constant, but small, abundance in the atmosphere. Because it is virtually chemically identical to ^{12}C (the higher mass creating only small differences), ^{14}C combines with oxygen to make heavy carbon dioxide, $^{14}C^{16}O_2$, and then finds its way into plants through photosynthesis, and thence through the food chain to the rest of the biological world. Living organisms continually exchange their carbon with the atmosphere via photosynthesis or respiration and food consumption, and live a short time, except for some very long-lived species of trees, compared with the half-life of ^{14}C. Thus, in the majority of living things, the ratio of ^{14}C to ^{12}C is a constant.

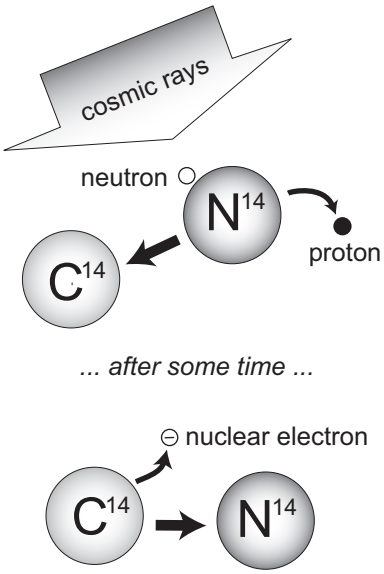


Figure 5.2 Production of ^{14}C from nitrogen and cosmic rays, and its decay. After Cloud (1988, p. 84).

When an organism dies, exchange stops, or actually slows, because bacterial and nonbiological processes still move materials in and out of the dead organisms, but at very low rates. The ^{14}C within the dead organism decreases over time according to the radioactive decay law. Biological materials that are less than roughly 60,000 years old have enough remaining ^{14}C that the electrons resulting from the decay can be directly counted in the laboratory, thus sensitively measuring the amount of ^{14}C remaining relative to the total carbon (masses 12, 13, and 14) in the sample. By comparing this number with the amount of ^{14}C relative to total carbon in the biosphere, and knowing the decay rate, the age is determined. The daughter, ^{14}N, is of no help, because it is identical to the rest of the ^{14}N that dominates our atmosphere and hence carries no signature of a radioactive origin.

In Chapter 21 ^{14}C dating figures prominently in the construction of a chronology of climate change in the latter part of the last ice age and the more recent, postglacial, period. It also has been a critical tool in dating ancient structures built, for example, in the arid southwestern United States by cultures now long gone the wooden beams and remains of fire pits being of key importance. Care must be taken, however, to understand the uncertainties that limit the accuracy of the age determinations made using ^{14}C dating.

In any scientific measurement, errors crop up associated both with the act of measurement and with the assumptions behind the interpretation. Measurement errors are familiar to anyone who has had to measure and build something. In using a ruler to determine what length of a beam to cut, for example, one might measure several times, or cut several beams to the same length. Repeated measurements or cuts reveal *random* errors, caused by small changes in positioning the ruler or the cutting tool. These errors generally can be estimated with some reliability, based on the sensitivity of the measurement and other factors. *Systematic* errors can be more insidious: in our example, perhaps the ruler is faulty, our carpenter has astigmatic eyesight, or the cutting

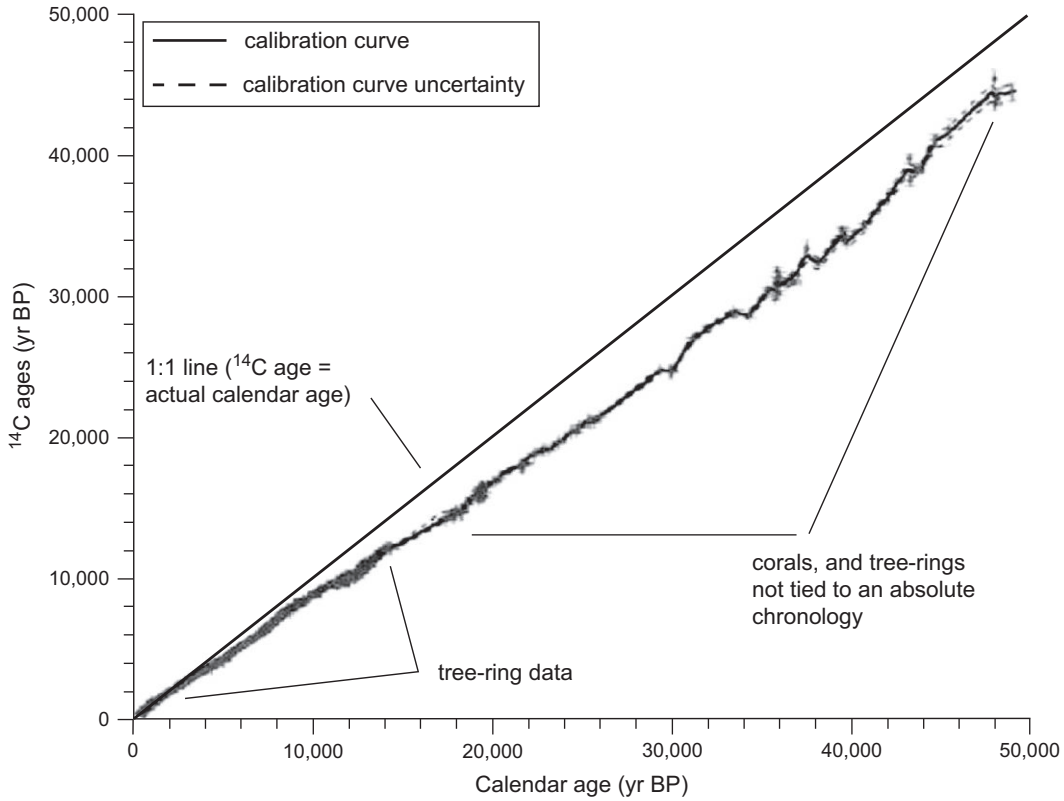


Figure 5.3 Relationship between actual age of a sample (cal yr before present; BP) and the age from ^{14}C dating (^{14}C yr BP). For the younger ages, tree-ring data (described in Chapter 21) are used; beyond about 10,000 years comparison between ^{14}C ages of corals and those derived from other isotopic systems is used, as well as some tree-ring data not tied to the more recent chronology. Modified from Fairbanks *et al*., 2005.

tool is out of alignment in some fashion. These errors can be difficult to detect but can ruin measurements. Experimental work in science is successful only so far as the errors can be reliably estimated and controlled.

In the interpretation of ^{14}C measurements, an obvious and crucial assumption has to do with the amount and rate of ^{14}C produced in the atmosphere over time. The manufacture of ^{14}C depends on the cosmic-ray flux, and this is known to vary as the strength of Earth's magnetic field changes, and as the Sun varies in its level of activity. Carbon-14 also may vary with changes in ocean circulation, which brings up varying amounts of carbon dioxide stored in deep water, or with other climatic or geologic events. The ages determined by ^{14}C dating must fold in these possible variations.

Cross-correlation, where possible, with independent dating techniques, for example, tree rings for more recent times (Chapter 21), is essential for calibration: it reveals that the ^{14}C level may have differed from recent values prior to about 3,500 years ago, necessitating a revision in some earlier dates from ^{14}C data. Tree-ring studies cannot go back over the tens of thousands of years accessible to ^{14}C dating, however, and so the earliest dates have larger uncertainties. Techniques such as independent dating of ages of corals can extend the calibration back to 50,000 years with somewhat less accuracy (Figure 5.3). Dates obtained with ^{14}C are generally younger than the actual age of the sample, and this discrepancy increases with age until it is several thousand years for ages of 20,000 years or more. Since the end of World War II, atmospheric nuclear testing has increased the production rate of ^{14}C in the atmosphere, so that the archaeologists of the future will need to correct for the increased amount of the isotope in organisms living at this time.

5.4 Measurement of parents and daughters: rubidium–strontium

The plausibility of dating very ancient events, such as the formation of the Earth and planets, by *radiogenic* (produced by radioactive decay) nuclides lies in the fact that these atoms are produced in stars in calculable amounts, expelled through supernova explosions, and decay in a regular fashion after formation. As elements, including the radioactive ones, became trapped in the solid material around our newly forming solar system, the initial abundances were modified through radioactive decay. The chemical affinity that particular elements have for certain rock phases provides the means for determining how much radioactive isotope was originally incorporated in the rock, and then the age since trapping via measurement of the present abundance of the radioactive isotope.

The main difficulty that confronts radioactive dating in which the decay process cannot be detected directly is the ambiguity

involved in knowing the actual initial abundances of the radioactive species and the decay products of the species. If 0.1 gram of the parent is present now in a sample, was there 1 gram when the rock formed? 2 grams? 10 grams? We could simply measure the amount of daughter product in the rock, but not all of the daughter product came from the decay of the parent. There might have been some daughter atoms present in the rock when it was formed. There is no guarantee that the rock was formed without any initial daughter atoms, and so, the only way to find the initial amount of daughter element is to find another isotope of the same element, not formed by radioactive decay, that acts as a chemical tracer of the daughter element.

Rubidium-87 (^{87}Rb) decays to strontium-87 (^{87}Sr) with a half-life of 49 billion years. There is a stable isotope of strontium, ^{86}Sr, formed directly in supernovas. Because isotopes of the same element behave nearly the same chemically, ^{86}Sr and ^{87}Sr would have tended to be trapped together in the same portions of the grains that form the rock under analysis. The ^{87}Rb would be locked elsewhere in the grains and, as it decayed, would have produced a local region of the rock enriched in ^{87}Sr.

Mathematically, the measured ratio of ^{87}Sr to ^{86}Sr in a particular grain equals the initial ratio at grain formation, plus the measured ratio of ^{87}Rb to ^{86}Sr multiplied by the number of mean-lifetimes that have elapsed. (This expression is approximate – the measured ratio of ^{87}Rb to ^{86}Sr is actually multiplied by $(e^{Rt} - 1)$. However, for the long half-life of ^{87}Rb our approximate expression suffices.) If we could find a part of our rock that contains no rubidium, that is, has only strontium, we could determine the initial strontium abundance and hence the age of the rock.

In real rocks, there is a range of strontium and rubidium abundances in different grains, and so we cannot isolate pure strontium. However, we can make a plot of the abundance of ^{87}Sr in each grain versus the abundance of ^{87}Rb, all relative to the ^{86}Sr abundance. Ideally, all grains in a rock of a given age should form a straight line on such a graph (Figure 5.4).

Where the straight line crosses (intercepts) the vertical axis is the initial amount of ^{87}Sr in the rock. This is the amount of ^{87}Sr that existed before the rock grains condensed from the primordial gas out of which the planets eventually formed. The age since the rock formed is just given by the slope of the line: the slope is the age divided by the mean lifetime of ^{87}Rb. Since the mean lifetime is measured in the lab, we can read the age in years from the slope of the plotted line. The older the rock, the less ^{87}Rb relative to ^{87}Sr there is in the sample today, and the steeper the slope of the line. The younger the rock, the more rubidium there is relative to strontium, and the less steep the slope.

Suppose that, after the rock formed, there was heating or chemical contamination that added some extra ^{87}Sr, leached out some ^{87}Sr, or did similar sorts of things to the ^{87}Rb. The affected grains would be skewed away from the ideal, straight line that represents a well-determined, single age since formation for the sample grains. Such a rock would not be suitable for dating, and this is immediately evident from the graph. The isochron technique therefore is self-correcting: a rock that has been severely altered will not have grains whose ^{87}Sr to ^{87}Rb analysis falls on a straight line (Figure 5.5).

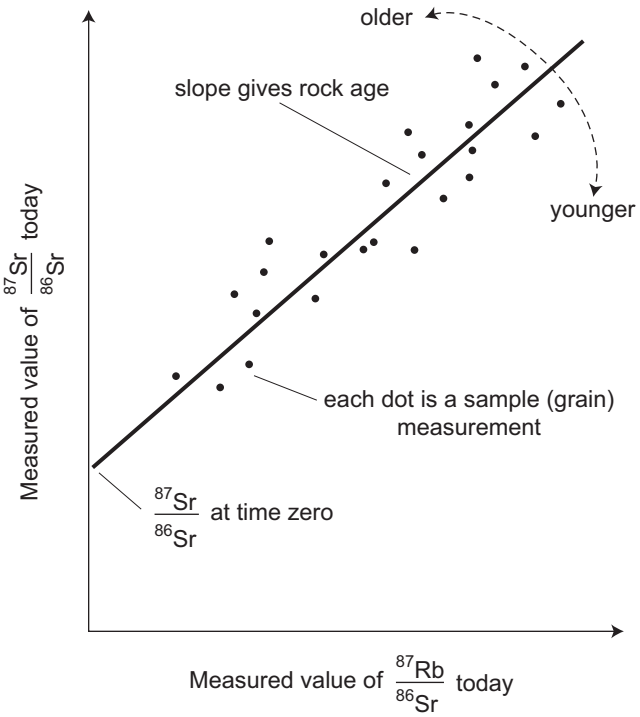


Figure 5.4 Amount of ^{87}Sr relative to the amount of ^{87}Rb at present, both shown as ratios to the ^{86}Sr abundance. Labels on the figure show how various quantities are determined.

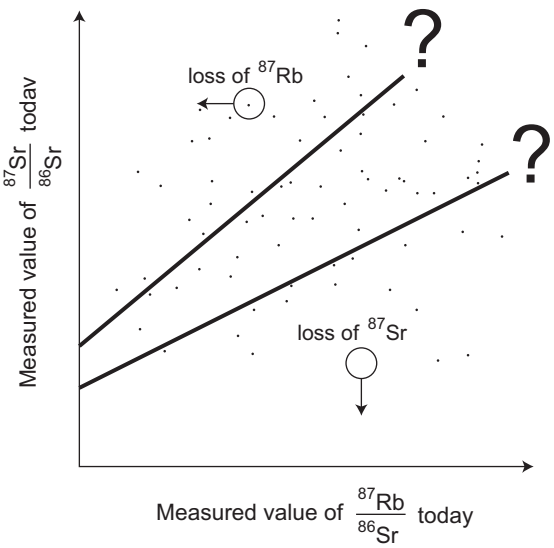


Figure 5.5 Possible effects on the age curve of chemical alteration of rock samples, in which either rubidium or strontium might be lost.

But are not all rocks altered to some extent? To determine the age of the solar system, we cannot use rocks that are now part of a planet – they likely have been melted at least once during and after the planet came together. Instead, the most primitive-appearing, least-altered meteorites are chosen to determine the age of the solar system. The meteorites with the most primitive

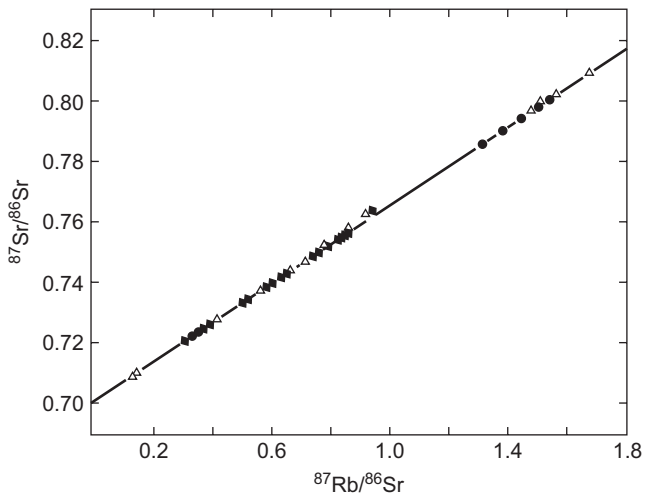


Figure 5.6 Determination of rubidium and strontium abundances for a number of meteorites whose appearance and chemical composition suggest that they have not been altered in planets. The data points fall very tightly on a curve that yields an age of 4.56 billion years. Modified from Minster *et al*. (1982) by permission of Macmillan Magazines Limited.

composition that we can find, which have an elemental composition as close to the Sun as possible, are the *carbonaceous chondrites*.

The results of an actual analysis on a suite of meteorites are shown in Figure 5.6. A number of isotopic parent–daughter systems are used to determine the age of the meteorites and solar system bodies; the pioneering work yielding essentially the currently accepted age was done by Claire Patterson of the California Institute of Technology in the 1950s, using the uranium–lead system. The ages of the most primitive meteorites center rather precisely on 4.56 billion years old. The oldest solids in the solar system, inclusions rich in calcium and aluminum embedded in these meteorites, cluster even more precisely around 4.568 billion years – that is, they all formed within a period of a million years early in the history of the solar system. The oldest Moon rocks brought back by the Apollo astronauts were found to be 4.4 billion to 4.5 billion years old using samarium–neodymium and lead isotopic systems. On Earth, one cannot find whole rocks older than 4.0 billion years, suggesting that no rocks in the upper parts of Earth escaped heavy alteration early in our planet's history. However, events earlier in Earth's history, such as formation of the core, can be determined using isotopic dating as described in Chapter 11.

Corroborating this determination are models of the structure of the Sun as a function of age. Models and observations of other stars show that a star expands slowly with time as it fuses hydrogen farther and farther from its central core. The radius and luminosity of the Sun correspond to the value expected at an age of roughly 4.6 billion years, consistent with the more accurate determination of the ages of the oldest meteorites.

5.5 Fission track dating

The dating techniques described above are only two examples of what can be done with radioactive isotopes. There are others. The spontaneous decay of heavy radioactive isotopes (mostly ^{238}U) can produce damage inside mineral crystals that shows up as micron-sized linear tracks. These "fission tracks" are produced at a rate that depends on the amount of ^{238}U. Thus if the uranium concentration of a sample is known, the density of tracks is an indication of the age of the sample. Particular minerals such as apatite and zircon are frequently dated using this technique. Ages ranging from centuries to hundreds of millions of years are accessible to fission track dating. The fading of tracks with time, at a rate dependent on temperature, is a potential complication, but this process can also be turned around to obtain the thermal history of a particular mineral.

5.6 Caveat emptor

This chapter has tried to explain simply and logically how scientists date geologic events and timescales using isotopes of elements that decay at measurable rates. These rates are determined in the laboratory and both theory and laboratory evidence point firmly to the concept that, even for half-lives of billions of years, the rates are constant and a function only of which particular isotope we are considering. Nonetheless, these techniques must be done carefully and require checks and cross-checks among different isotopic systems and painstaking error analysis.

The age determinations are not easy, but the reproducibility that has been achieved from sample to sample and system to system makes the determinations of cosmic ages highly convincing. Particularly firm is the age of the solar system as defined by the appearance of the first planet-building solids, and hence the beginning of the formation of Earth and its planetary neighbors, as 4.568 billion years ago. Radioactive isotopes have been used to estimate the time since element formation began in the stars. This timescale is much less certain than the age of the solar system, but still a useful gross constraint on the age of the universe, which is independent of the galactic red-shift argument. Isotopic systems are also the foundation for dating geologic events on Earth itself, forming the backbone of the history described in Part III.

Summary

Determining the time when events occurred during the history of the Earth and the solar system, or "dating" of events, is done in two ways. A *relative* chronology is one in which a series of events is ordered into a temporal sequence, from oldest to youngest, without fixing the specific times they occurred in history, or the duration of time between events. Determining the relative ages of sedimentary layers on the Earth, or which terrains on the Moon are the oldest based on the number of impact craters, are examples. An *absolute* chronology contains information on the actual time in history at which an event occurred, or its duration. Providing an absolute date for a geologic sample requires the presence of natural clocks that can be calibrated and whose uncertainties can be constrained by cross-checking among samples or different techniques. Radioactive isotopes are natural clocks because the rate of decay of a given isotope of an element can be measured accurately in the laboratory. Different radioactive isotopes decay at vastly different rates, allowing diverse types of events to be dated. Carbon-14 dating of biological samples – usually pieces of wood – allows the time since the organism died to be determined, up to ages of tens of thousands of years. The technique relies on the continued circulation of carbon from Earth's atmosphere through organisms as long as they are alive, and is accurate to perhaps 10% in the age, based on comparison with other age determinations. For the geologic history of the Earth and solar system, up to and including the age of the Earth, isotopes that decay much more slowly, and that tend to accumulate in rocky material, must be employed. Because a given rock sample may have contained the decay product of the radioisotopic clock when it formed, usually age determinations require measuring the amount of the decaying "parent" isotope and its "daughter" product relative to a third, stable isotope. In this way it is possible to see whether multiple episodes of melting or other alteration of the rock may have degraded the record of its age. Several isotopic systems with slow decay rate have been used to date with high precision the appearance of the first solids in the solar system, the age of the Moon, and key events in the history of the Earth.

Questions

1. What possible sources of contamination might a geochemist have to guard against while radioisotopically dating rock samples?
2. What arguments would you make to justify the assertion that the radioactive decay rates of the elements have not changed drastically over the age of the universe? Hint: consider the determination of the beginning of element formation by stellar nucleosynthesis.
3. The most precise determinations of the ages of meteorites are made with isotopic systems other than rubidium–strontium. With a short literature search, find information on how the ages are determined from these other isotopic systems.
4. Radioisotopic dating of samples is done in laboratories on Earth, but proposals recently have been made for miniature labs on Mars to do remote dating. Given the limited sensitivity of miniature laboratories, what isotopic parent–daughter pair would you choose from Table 5.1 to ensure a high abundance in the rocks on Mars? (Look ahead to Chapter 15 for information on Mars).

General reading

Broecker, W. 1985. *How to Build a Habitable Planet*. Eldigio Press, New York.

References

Allègre, C. J., Manhès, G., and Göpel, C. 1995. The age of the Earth. *Geochimica et Cosmochimica. Acta* **59**, 1445–56.

Bouvier, A. and Wadhwa, M. 2010. The age of the Solar System redefined by the oldest Pb-Pb age of a meteoritic inclusion. *Nature Geosciences* **3**, 637–41.

Cameron, A. G. W. 1993. Nucleosynthesis and star formation. In *Protostars and Planets III* (E. H. Levy and J. I. Lunine, eds). University of Arizona Press, Tucson, pp. 47–73.

Cloud, P. 1988. *Oasis in Space: Earth History from the Beginning*. W. W. Norton, New York.

Fairbanks, R.G., Mortlock, R. A., Chiu, T.-C. *et al*. 2005. Radiocarbon calibration curve spanning 0 to 50,000 years BP based on paired ^{230}Th/ ^{234}U/^{238}U and ^{14}C dates on pristine corals. *Quaternary Science Reviews* **24**,1781–96.

Minster, J. F., Birck, J.-L., and Allègre, C. J. 1982. Absolute ages of formation of chondrules studied by the ^{87}Rb-^{87}Sr method. *Nature* **300**, 414–19.

Patterson, C. 1956. Age of meteorites and the Earth. *Geochimica et Cosmochimica Acta* **10**, 230–7.

Considine, D. M. (ed.) 1983. Radioactivity and other dating techniques. In *Van Nostrand's Scientific Encyclopedia*. Van Nostrand Reinhold, New York, pp. 2387–9.

Swindle, T. D. 1993. Extinct radionuclides and evolutionary timescales. In *Protostars and Planets III* (E. H. Levy and J. I. Lunine, eds). University of Arizona Press, Tucson, pp. 867–81.

Wagner, G. A. and Van den haute, P. 1992. *Fission Track-Dating*. Kluwer Academic Publishers, Dordrecht.

6

Other uses of isotopes for Earth history

Introduction

In addition to the dating of rocks by measuring amounts of radioactive isotopes and their decay products, isotopes can be useful as indicators of climate variations on Earth over its long history. Here, the key is to use stable isotopes of the same element. The difference in mass between the isotopes leads to separation, called fractionation, of the isotopes in natural systems; the separations in some cases are a function of the climate, specifically temperature.

To use isotopes as climate indicators, four key features are required:

1. availability of stable isotopes of the same element whose separation depends on temperature
2. incorporation of the fractionated isotope mixture in some storage medium that is preserved for a long time
3. ability to measure accurately the ratio of the various isotopes
4. a means to date, in an absolute or a relative sense, the age of the stored isotope data.

6.1 Stable isotopes, seafloor sediments, and climate

6.1.1 Carbon

Three important elements for tracking climate changes are carbon, oxygen, and hydrogen. Consider the carbon first. Carbon has two stable isotopes, ^{13}C and ^{12}C. Recall that ^{14}C is radioactive and used for dating relatively recent events. Certain biological processes distinguish mass differences in isotopes. We cannot survive on deuterated water (HDO or $^{1}H^{2}HO$). Likewise, plants are observed to preferentially take up ^{12}C in carbon dioxide (CO_2), and hence preferentially enrich the atmosphere in ^{13}C. The more temperate the climate, the more land area that is available for plants, and the more ^{12}C that is taken up. In ice ages, global plant activity is reduced, and so less ^{12}C is taken up.

How is the excess or deficit of ^{13}C in the atmosphere recorded? Many single-celled sea organisms produce protective shells or plates, for example the "coccolithophorids". The plates or *coccoliths* are composed of calcite, $CaCO_3$, where Ca is the element calcium. The carbon in the calcite comes from the carbon dioxide in the atmosphere, and the organisms are observed to take up ^{13}C and ^{12}C with equal propensity. Therefore, coccoliths of tiny sea organisms record the atmospheric ratio of ^{13}C to ^{12}C. When the organisms die, the shells fall to the ocean floor, where they are buried over time. The calcite shells are a biologically produced mineral that can remain intact on the seafloor for enormous lengths of time.

Ancient seafloor sediments often have been uplifted in more recent mountain chains through geologic processes. As the ancient sediments, now hardened as rock, are raised and then exposed to view by erosion, the calcite contained within can be analyzed for the ^{13}C to ^{12}C ratio. The higher the ^{13}C amount, the higher the productivity of the biosphere was at the time the shell-forming creature lived. Although a number of factors affect productivity, one such is the global mean temperature. If the sediments can be dated, by means of radioactive isotopes or by relative techniques involving fossils (Chapter 8), information on the climate as a function of time thus is determined through this proxy measure.

6.1.2 Oxygen

The oxygen bound up as water in the oceans also produces a record of Earth's surface temperatures. Water on Earth is a mixture of the stable isotopes $H_2{}^{16}O$, $H_2{}^{17}O$, and $H_2{}^{18}O$, with the ^{16}O being by far the more abundant. From laboratory measurements, $H_2{}^{16}O$ is known to be preferentially evaporated from the

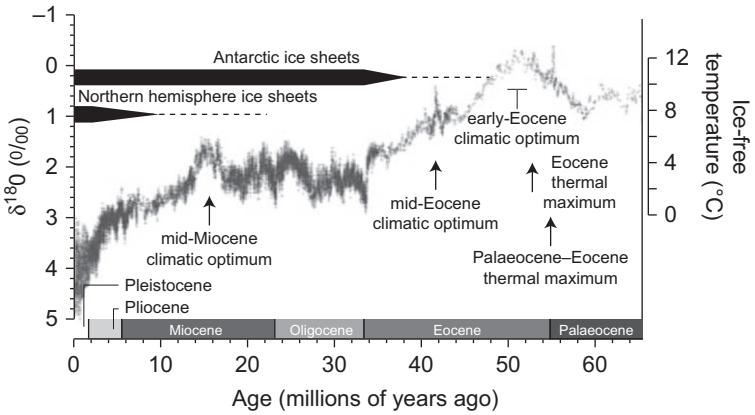


Figure 6.1 Record of ^{18}O enhancements from fossils of benthic foraminifera, ocean creatures in seafloor sediments, ranging in age from 65 million years ago to the present. The horizontal axis is time before present; the corresponding geological epochs of the Cenozoic era (Chapter 8) are marked on the figure. The measure of isotopic enrichment, $\delta^{18}O$, is by convention the difference of $^{18}O/^{16}O$ in the sample to that of a known standard, divided by that reference standard $^{18}O/^{16}O$ value. As is common to avoid having to use decimals and zeros, the values are all multiplied by 1,000. Larger amounts of $\delta^{18}O$ correspond to larger global ice volume and hence colder conditions, and a rough temperature calibration is shown on the right-hand axis down to the freezing point of water. Major climate events over this time period are labeled, as are the onsets and durations of the large Antarctic and northern hemisphere ice sheets. The geologic epochs along the bottom are explained in Chapter 8. Adapted from Pälike and Hilgen, 2008.

oceans to form clouds. During cold periods on earth – ice ages – the polar caps were built up and spread outward in the form of ice sheets. The sources of the ice sheets are storm systems that dump large amounts of moisture on high-latitude continents in the form of snow. Therefore, water actually was lost from the oceans during ice ages and stored as ice in great continental ice sheets. Because it is the lightest isotope $H_2{}^{16}O$ that is preferentially evaporated from oceans, during ice ages the oceans should be enriched in ^{18}O. The same kind of enrichment occurs for ^{17}O, but its mass difference from ^{16}O is half that of ^{18}O, and so, climate effects are smaller. For that reason, we focus only on the $^{18}O/^{16}O$ record because it is read more easily.

Again, a storage medium for the oxygen is required and, as before, tiny sea creatures play the role. To make the calcite shells or plates such as coccoliths requires oxygen as well as carbon ($CaCO_3$). As with carbon, the oxygen is taken up without regard for the isotope number, ^{18}O and ^{16}O equally. The oxygen source is water in the ocean. Hence these microscopic shells record the ratio of ^{18}O to ^{16}O in the ocean at the time of their formation. The measurement for oxygen (and carbon, for that matter) is available from organisms at both the ocean's surface and its depths, because oxygen extracted by deep-sea shellmakers will record the local $^{18}O/^{16}O$ ratio in the ocean water.

As the organisms die, the calcite shells are deposited on the ocean floor, buried by progressive sedimentation and become part of the rock record. Ancient ocean sediments exposed by geologic events allow the calcitic oxygen to be extracted, the $^{18}O/^{16}O$ ratio measured, and the global temperature determined. To put these data into a climate chronological sequence, the sediments or surrounding rock layers then must be dated. Figure 6.1 shows an example of ocean temperatures derived from the oxygen isotopic data for the time period from 65 million years ago, which covers the demise of the dinosaurs through the golden age of mammals (see Chapter 19). It should be remembered that the oxygen isotopic ratio is only an imperfect measure of ocean temperature, since other effects such as ocean salinity can alter isotopic ratios. Hence cross-comparison with other types of data, such as distribution of planet species in the fossil record, is essential.

6.1.3 Hydrogen

About 10 parts per million (ppm) of water on Earth contains heavy hydrogen, or deuterium, primarily paired with a light hydrogen to make HDO (as opposed to normal water, H_2O). As with the heavier isotopes of oxygen, deuterated water tends to be preferentially left behind during evaporation of ocean water near the equator. Thus air masses moving away from the equator and hence toward colder latitudes are slightly enriched in normal water; that is, they possess somewhat less than the 10 ppm of HDO that is typical for the ocean. As rain and snow form in the air mass, the deuterated water is preferentially and progressively removed from the storm system in the precipitation, so that storms near the poles drop snow that is significantly depleted in deuterium. The depletion is exaggerated during colder climate episodes relative to warmer, because the drop in temperature from equator to pole is larger during colder times. (This is checked by mapping the distribution of plant species during warm times versus ice ages.) Also, the tendency of deuterium to fall out in the rain and snow is exaggerated at lower temperatures (Figure 6.2).

The resulting *deuterium fractionation* has been used to study the most recent epochs of glacial climate and warm episodes in between, by sampling the ice sheets that cover Antarctica, Greenland, and other very cold places. The record in such ice sheets extends back less than 300,000 years in Earth history, but it is much more detailed than that in the more ancient seafloor sediment record of carbon and oxygen isotopes. Colder times are characterized by more exaggerated deuterium fractionation and hence greater deuterium depletion in the ice laid down at that time. Warmer episodes show less deuterium depletion. Because the ice also contains the oxygen isotopes discussed above, the

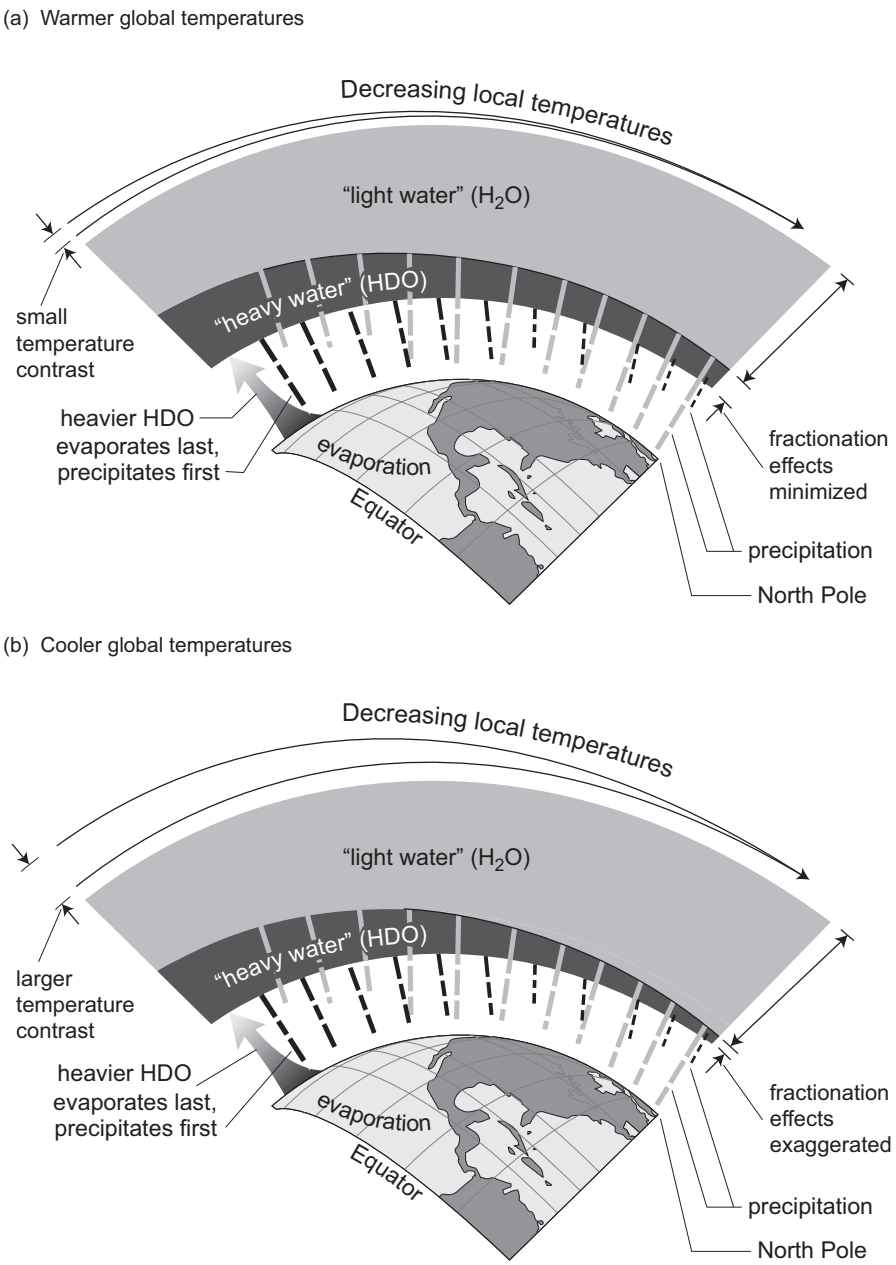


Figure 6.2 Progressive depletion of deuterium-bearing water from equatorial ocean to the polar ice sheets, during (a) warmer and (b) colder times. The steeper equator-to-pole temperature drop under cold climate conditions exaggerates the fractionation relative to that in warmer times. HDO is shown in black, H_2O in gray.

deuterium and ^{18}O depletions can be compared to help build confidence in the paleo-temperatures (ancient temperatures) derived from the core. Chapter 21 discusses the application of stable isotopes to understanding the nature of interglacial warm periods such as that in which we now live.

6.1.4 Other systems

There are many other methods for determining past temperatures on Earth; here we have described only a few widely applied techniques. Nitrogen isotopic ratios are a good indicator of phytoplankton productivity, and as well sulfur isotopes provide a range of information on the state of past environments on Earth. Also, certain organic molecules derived from algae are composed of long chains of atoms that preserve well in sea sediments. The relative abundances of some of these molecules, called *alkenones*, reflect sea surface temperature at the time of their formation. The advantage of this so-called *UK 37 index* is that it is not influenced by salinity to the same extent as the oxygen isotopes.

6.2 A possible temperature history of Earth from cherts

Oxygen isotopic exchange potentially provides a temperature indicator back through 80% of Earth's history. *Cherts*

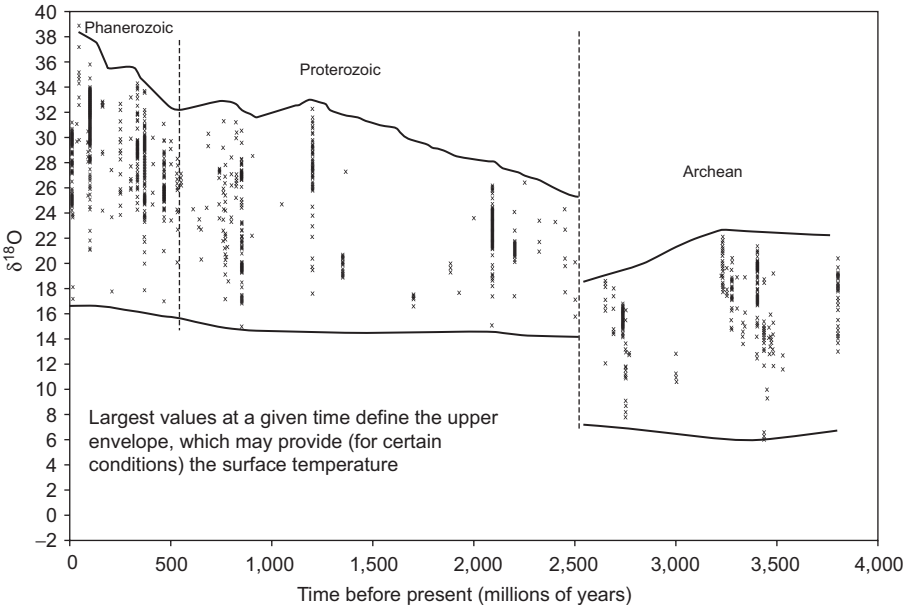


Figure 6.3 A possible temperature history of the Earth is reflected in the oxygen isotopic ratio in samples of chert. Plotted against time in Earth's history is the so-called $\delta^{18}O$, which is the difference of the ratio $^{18}O/^{16}O$ measured in the sample to that in a reference standard, normalized by the reference standard. As is common to avoid having to use decimals and zeros, the values are all multiplied by 1,000. As discussed in the text, an estimate for the surface temperature of the Earth is given by the upper envelope of the curves. Geologic era, which we introduce in Chapter 8, are labeled and their boundaries indicated by dashed lines. From Knauth (2005).

are hard rocks, composed largely of very-fine-grained silica. Silica is formed from silicon and oxygen: SiO_2. It can occur as bands in limestones, as nodules, or in other physical forms. Cherts form in a wide range of environments, precipitating directly out of rivers or ocean waters, or forming from rocks that are subjected to mild increases in temperature and pressure. *Biogenic* chert, that is, chert made by organisms such as sponges or radiolaria that secrete silica, is probably the most abundant.

Of interest to us here is that the oxygen isotopic content of the chert bears a definite relationship to that of the environment in which it is made. If precipitation occurs in an ocean environment, then the ^{18}O content of the silica decreases with increasing temperature in a manner that can be quantified in the laboratory. Essentially, the chert, which preserves very well as a sediment through time, acts to record the ambient water temperature through the oxygen isotopic enhancement during its formation. High temperatures in the water from which the chert is precipitated lead to lower $^{18}O/^{16}O$ ratios in the chert, while low temperatures lead to high $^{18}O/^{16}O$ ratios.

Unfortunately, using cherts as indicators of the surface temperature of Earth is extremely complicated because cherts form in so many different environments and the $^{18}O/^{16}O$ values may be altered in ways that have nothing to do with the surface temperature. In the 1970s geochemists Paul Knauth at Arizona State University and Donald Lowe at Louisiana State University attempted to use cherts to determine ancient ocean temperatures in spite of these difficulties. They argued that, for most (but not all) types of chert, processes during or after formation would tend to lower the $^{18}O/^{16}O$ enhancement in cherts relative to the value obtained during precipitation from ocean waters. Therefore, for a collection of cherts of a given age, the cherts with the highest $^{18}O/^{16}O$ values should most nearly reflect equilibration with ocean waters during formation. Hence, the cherts with the highest $^{18}O/^{16}O$ value at a given time provide a measure of Earth's ocean temperature.

A relative temperature history of the Earth from cherts is shown in Figure 6.3 from Knauth (2005). If one calibrates the oxygen isotopic data with the recent temperatures of 10 to 15 °C, the drop in $^{18}O/^{16}O$ going back in time implies a temperature of 55 to 85 °C in the first third of the Earth's history. Particularly striking is the sharp change in global temperature at about 2.5 billion years ago. This sharp drop in temperature occurs roughly in the time range where other types of geologic evidence suggest that the Earth experienced at least two episodes of dramatic global cooling in which ice covered much of the Earth ("snowball Earth"; see Chapter 19). Also, as Knauth (2005) points out, the earliest life forms, based on study of the relationships in the genome of organisms existing today (Chapter 12), preferred environments as warm as those implied by the earliest chert data.

Some qualifications must be applied to this analysis. First, the chert samples were formed over a range of latitudes, leading to the concern that one is mixing latitudinal and time variations in temperature. As we discuss in Chapter 19, however, warmer ice-free climates, which have dominated over most of Earth's history, experienced much less variation of temperature with latitude than we experience today. The second issue is more serious, and has to do with whether the oceanic value of $^{18}O/^{16}O$ has really been constant over time. One source of variation are the episodes of massive glaciation interspersed throughout Earth's

history that are mentioned above. Formation of glaciers alters the baseline $^{18}O/^{16}O$ value in the oceans. Further, the baseline $^{18}O/^{16}O$ abundance may have been lower than today's during the first quarter of Earth's history, based on the chemistry of the most ancient cherts, which would explain the discontinuity in $^{18}O/^{16}O$ at 2.5 billion years ago without invoking a sharp drop in temperature. Finally, one must remember that the surface temperature is interpreted to be provided by the chert samples with the highest $^{18}O/^{16}O$ out of a large range. There is no guarantee that the uppermost value actually corresponds to the ambient surface temperature, even though this is a reasonable assumption for the later data. If areas of high geothermal activity were more prevalent in the ancient past than today, it is more likely that the older samples, even those with the largest $^{18}O/^{16}O$, are affected by, or indeed predominantly reflect the environment of, these hot spots. That many of the samples reflect conditions in and near hot spots, such as hydrothermal vents on the ocean floor, is supported by measurement of silicon isotopic ratios, which vary in the chert samples over time in a way that suggests alteration in hot vents.

The chert story illustrates how important it is to be cautious and skeptical when extracting conclusions from a single type of data. Unfortunately, other evidence for Earth's temperatures in the first quarter of its history is extremely sketchy. The intriguing conclusion from the chert data, that at least portions of the Earth's oceans were as warm if not warmer than today, is consistent, at least, with the extensive geologic evidence that liquid water was stable on Earth at least 3.8 billion years ago. Although this may not seem surprising, it is of some significance because stellar models argue that the Sun was much less luminous at that time than it is today. The geologic record tells us that this early warm period was punctuated (or even terminated) by glacial epochs, in which ice covered much or perhaps all of the Earth, beginning about 2.9 billion years ago. Evidently the Earth's climate underwent an adjustment that we do not understand, or periods dominated by ice existed earlier (but were not recorded in the chert data); however, the geologic evidence is too scant to record this.

We discuss the *faint early Sun* problem in Chapter 14. For now, it suffices to note that the interpretation of the oxygen isotopes in chert as indicating high surface temperatures on the early Earth create a paradox because they require early Earth to have been significantly warmer than at present when, in fact, the Sun was significantly dimmer. Spacecraft images of Mars and direct measurement of rocks at the Martian surface also indicate that our neighboring planet was hotter earlier in its history than it is today.

This dual-planet dilemma regarding dramatic climate change early in the history of the planets, in the face of the faintness of the Sun at the time, represents a major puzzle that we must tackle later in the book.

Summary

Stable isotopes of major elements, that is, isotopes that do not decay measurably over Earth's history, can be used to track the climate history of our planet. In order to use stable isotopes, the given element must be commonly present in sediments or life forms, it must have more than one stable isotope whose separation depends on temperature, the altered isotopic ratio must be preserved in a time-ordered or datable way for a long time, and the isotopic ratios must be measurable. For recent climate, isotopes of carbon, oxygen, and hydrogen are available. Carbon has two stable isotopes, of mass 12 and 13, respectively, and the lighter isotope is preferentially incorporated from atmospheric carbon dioxide into carbohydrates produced in plants by photosynthesis. Thus, during warm periods, when less land is covered by ice and more rainfall occurs, allowing more plant activity, the ratio of the heavier to the lighter isotope, ^{13}C to ^{12}C, is enriched in the atmospheric carbon dioxide. This record is preserved by shell-forming organisms that take up the atmospheric carbon for their shells, and upon dying become part of the sediments on the ocean floor. Oxygen and hydrogen isotopes record climate based on the difference in propensity for evaporation between the lighter and heavier isotopes. Because the temperature drops more steeply from equator to pole during cold periods relative to warm ones, preferentially more of the lighter isotope is extracted from the ocean water in cold times and sequestered at the poles as ice. Thus, in colder times the enrichment of the heavier isotope in the ocean water is larger than in warmer times. For oxygen this is recorded in shells; for hydrogen the record is in the cores of ice deposited during colder climates and preserved in Antarctica and elsewhere. An oxygen isotopic record of ancient climate exists in silicon–oxygen rocks called cherts. The cherts form by precipitation from ocean water, and the isotopic ratio of oxygen is altered as a function of temperature during the precipitation into the mineral phase. The cherts suggest that temperatures in the ocean were higher in the early history of the Earth than is the case today; however, the chert record may be contaminated by a number of factors other than the mean ocean temperature, including the effect of high-temperature "hydrothermal" vents.

Questions

1. Why is it important to use more than one isotopic system to determine the history of Earth's surface temperature?
2. What is it about the possible differences between ^{12}C and ^{13}C that would lead to a preferential uptake by planets of the former compared to the latter? Is this a phenomenon of chemistry? If so, could abiotic chemical processes in, for example, the atmospheres of Mars (Chapter 15) or Saturn's moon Titan (Chapter 16) exhibit the same sort of fractionation of the isotopes? Or is this a possible way to detect biological versus purely abiotic chemical processes?
3. The carbon and oxygen isotopic ratios in shell-forming organisms is not entirely independent of oceanic conditions; rather, these ratios might be altered by the amount of hydrogen carbonate ions (ions with the formula HCO_3^- historically called bicarbonates) in the oceans (see Chapter 14 for a discussion of the chemistry). As the carbonate ions (which have the formula CO_3^-) become more abundant in the ocean, the values of $^{13}C/^{12}C$ and $^{18}O/^{16}O$ decrease in the shells. Since hydrogen carbonate ions are primarily produced from erosion of rock by rainfall, and then ends up in the oceans by river runoff, what might you predict would be the direction of this effect given that rainfall is decreased during colder epochs? How might you correct for this effect in determining past ocean temperatures from the shells of organisms?
4. In using cherts to determine global temperatures in the past, how would you test the claim that the oceanic ^{18}O value has been constant over Earth's history?

General reading

Considine, D. M. (ed.) 1983. Cherts. In *Van Nostrand's Scientific Encyclopedia*. Van Nostrand Reinhold, New York, p. 624.

Kasting, J. F. and Kirschvink, J. 2012. Evolution of a habitable planet. In *Frontiers of Astrobiology* ed. C. Impey, J. Lunine and J. Funes. Cambridge University Press, in press.

References

Jouzel, J. and Merlivat, L. 1984. Deuterium and oxygen 18 in precipitation: modeling of the isotopic effects during snow formation. *Journal of Geophysical Research* **89**, 11,749–57.

Knauth L. P. 2005. Temperature and salinity history of the Precambrian ocean: implications for the course of microbial evolution. *Palaeogeography, Palaeoclimatology, Palaeoecology* **219**, 53–69.

Knauth, L. P. and Lowe, D. R. 1978. Oxygen isotope geochemistry of cherts from the Onverwachte group (3.4 billion years), Transvaal, South Africa, with implications for secular variations in the isotopic composition of cherts. *Earth and Planetary Science Letters* **41**, 209–22.

Pälike, H. and Hilgen, F. 2008. Rock clock synchronization. *Nature Geoscience* **1**, 282.

Prahl, F. G. and Wakeham, S. G. 1987. Calibration of long-chain alkenones as indicators of paleoceaongraphic conditions. *Nature* **330**, 367–9.

Shackleton, N. J. 1986. Paleogene stable isotope events. *Paleogeography, Paleoclimatology, Paleoecology* **57**, 91–102.

Spero, H. J., Bijma, J., Lea, D. W., and Bemis, B. E. 1997. Effect of seawater carbonate concentration on foraminiferal carbon and oxygen isotopes. *Nature* **390**, 497–500.

Van den Boorn, S. H. J. M., van Bergen, M. J., Nijman, W., and Vroon, P. Z. 2007. Dual role of seawater and hydrothermal fluids in Early Archean chert formation: evidence from silicon isotopes. *Geology* **35**, 939–42.

Vostok Project Members. 1995. International effort helps decipher mysteries of paleoclimate from Antarctic ice cores. *EOS* **76**, 169.

7

Relative age dating of cosmic and terrestrial events: the cratering record

Introduction

The absolute dating techniques of Chapter 5 rely on very precise laboratory analyses of rock samples. For Earth, an abundance of accessible samples exists. However, with respect to the rest of the solar system, only meteorites, small bits of asteroidal and cometary debris – interplanetary dust particles (IDP), and samples from the Moon have been delivered to terrestrial laboratories for age analyses. One class of meteorites, the Shergottites–Nakhlites–Chassigny (SNC), may have been ejected from Mars by collision with one or several asteroids. Aside from these cases, we have no known samples of material from large bodies in the solar system and thus cannot date major geologic events on the surfaces of the bodies in an absolute fashion.

Instead, scientists use *relative* dating techniques to infer time histories of the moons and planets in the solar system, and they rely primarily on the record of bombardment, or *cratering*, of the surfaces of these bodies. We describe this technique and the physics of cratering in the present chapter. In addition to providing a foundation for inferring key aspects of the solar system's history, this discussion provides a good foundation for the presentation in Chapter 8 of relative age dating on Earth, which relies on geologic processes other than cratering but for which the principles are much the same.

7.1 Process of impact cratering

Impact cratering is a process in which a high-speed projectile collides with a solid surface, forming an excavated region called a crater. Impact craters, and the closely related form of craters caused by massive explosions, such as nuclear detonations, can be distinguished from those produced by other processes, such as volcanism or collapse due to groundwater withdrawal, by their distinctive appearance (Figure 7.1).

Projectiles impact with velocities imparted by virtue of their orbital motion and the gravitational pull of the target planet. Impact speeds vary depending on the target planet's distance from the Sun and the strength of its gravitational field. Typical impact speeds onto the Moon, due to the free-fall velocity of projectiles at 150 million kilometers from the Sun, are 40 kilometers per second, or just short of 100,000 miles per hour.

An automobile hitting a surface at 160,000 km (approx. 100,000 miles) per hour delivers an impact energy a million times higher than if it were in a head-on collision with another vehicle at 80 km (approx. 50 miles) per hour, that is, 160 km (approx. 100 miles) per hour relative velocity, because impact energy scales as the square of the velocity. However, it also scales with the mass, and the bigger craters on planetary surfaces are formed by impactors that are kilometers in size. The energy released by just one such impactor, kilometers across, is equivalent to the release of the world's entire nuclear arsenal – at the peak prior to current disarmament – many dozens of times over! Such an enormous release of energy on a habitable planet has the capability to transform oceans and atmospheres, and to destroy life on a planetary scale.

At impact with the ground, the projectile plows into the surface, its energy of motion rapidly converted into heat, and the impact itself sends *shock* waves, familiar examples of which are thunder and sonic booms, into the ground. The ground itself is compressed and shattered by the enormous temperatures and pressures of the shock wave. The projectile also is shocked and shattered. The shock waves travel outward and downward in the ground in a hemispherical pattern. Nearest the impact, rock is vaporized or melted; farther away it is pulverized. As the shock waves travel away from the impact, the ground begins to rebound toward the center of the hemispherical cavity or crater, forming a central peak in the case of moderate-sized to large impact craters. The central peak can form only because the rock is in a temporary state of being partly molten and partly solid,

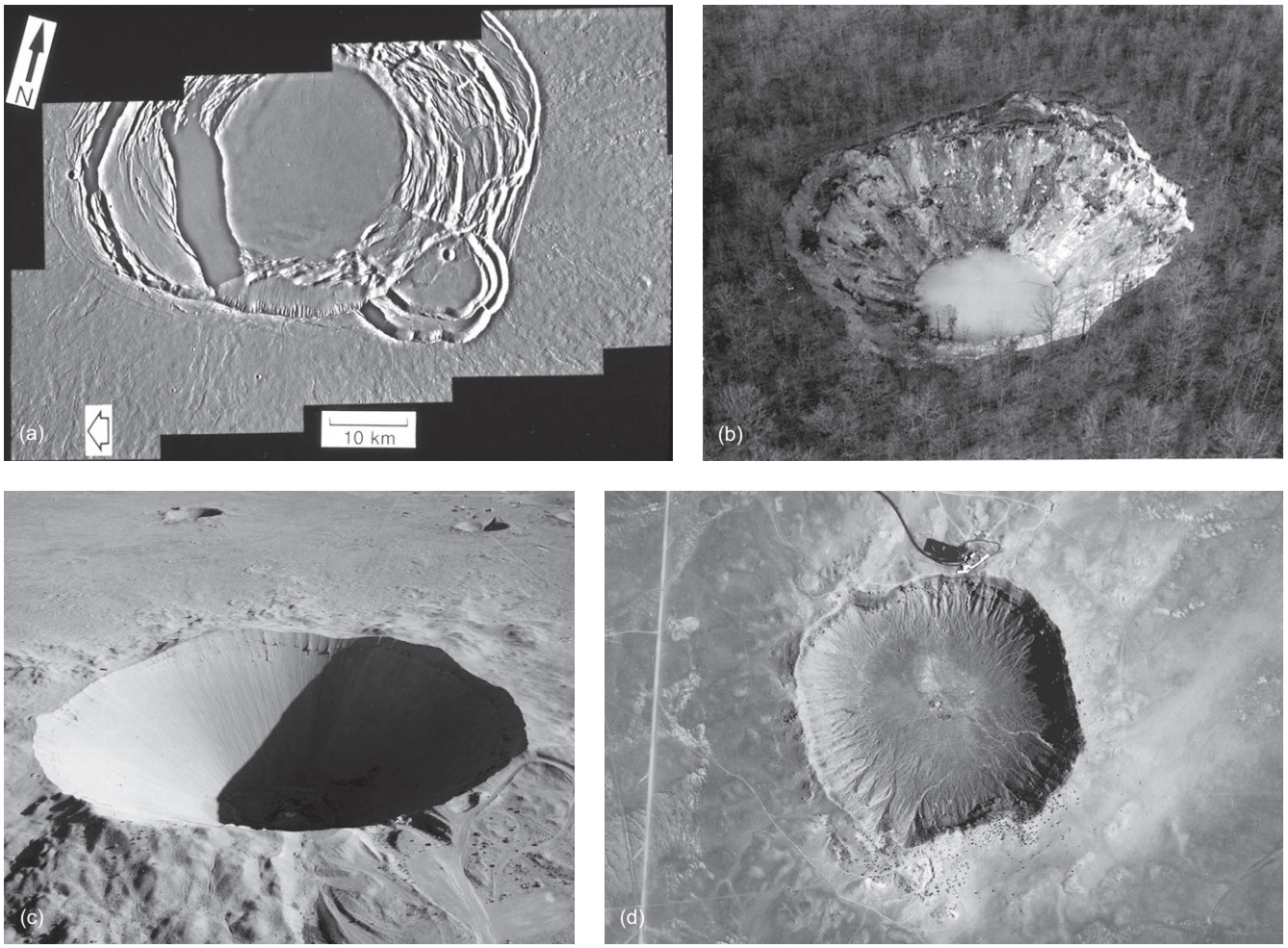


Figure 7.1 Examples of craters formed by different processes: (a) Caldera at the summit of the Martian volcano Aescraeus Mons (mosaic of NASA *Viking* images generated by J. Zimbelman at the Lunar and Planetary Institute); (b) sinkhole near Montevallo, Alabama, 120 meters across, formed by the action of groundwater (US Geological Survey photo); (c) explosion crater about 250 meters across in Nevada, generated by a 30-kiloton nuclear warhead detonated underground (US Department of Energy); (d) meteor crater, Arizona, a small (1 km diameter) impact crater.

the solid part being so weak that the shock waves moving back and forth can readily push the material. As the shock waves dissipate, the central peak remains intact.

The shock waves also raise a rim around the crater as well as eject material off the sides, this material (ejecta) shooting into the air as hot molten (liquid) rock, traveling many times the size of the crater away from the center, forming lines of smaller *secondary* craters as well as streaks or *rays* of material as it strikes the ground. The impactor itself is obliterated and becomes a small part of the ejecta.

Figure 7.2 shows the stages of crater formation and the final shapes of typical small and large craters. There are many variations: small craters do not have well-developed central peaks. Extremely large impactors send shock waves through deeper parts of the target's interior, where the warmer rock or ice can flow more easily, creating large-scale wave patterns that are preserved as *multiring basins*. Mare Oriental on the surface of Earth's Moon, Valhalla on Jupiter's moon Callisto, and Gilgamesh on Jupiter's moon Ganymede are examples. Craters also may take on different forms depending on whether ground ice is present, and the strength of the planetary crust: weak crusts will cause crater topography to disappear over time, leaving only ghostly outlines. Finally, erosion by water and subduction of crust (Chapter 9) have removed most of the craters on Earth, and left many others barely discernible (Figures 7.3a–f).

7.2 Using craters to date planetary surfaces

Craters can be used to determine how old one surface is relative to another because the rate of impacts over time is thought to have declined slowly over the past three-quarters of solar system history, having decreased quickly prior to that from a much larger initial rate. Surfaces that are young, that is, which have been renewed through lava flows, mountain building, erosion by water, and other geologic processes, will show fewer craters than surfaces that are much less active, or older. Because of this we can use the abundance of craters on various surfaces of a planetary body to determine, in a relative sense, when certain kinds of geologic processes occurred relative to others. The

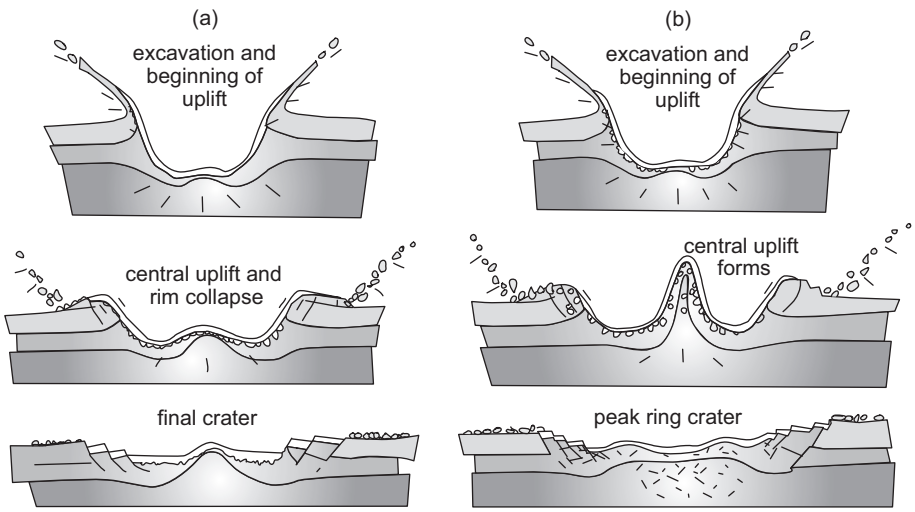


Figure 7.2 Stages in the formation of (a) small and (b) large craters. Modified from Melosh (1989, p. 142) by permission of Oxford University Press.

freshness of craters, that is, how bright their debris or ejecta blankets appear and how sharp their features are, provides an additional refinement to the relative dating process. It is also possible to compare relative ages from one planet or moon to another, provided one can calibrate the rate of impacts in one part of the solar system relative to another.

An example from the Moon provides a classic illustration. The bright areas of the surface of the Moon are very heavily cratered regions called the *lunar highlands*, as revealed by telescopes and images from lunar orbiting spacecraft (Figure 7.4a). In these regions the density of craters is so great that craters overlap with and are superimposed on each other; down to the limit of resolution on the image, one sees a scene filled with craters.

The dark portions of the Moon, on the other hand, consist of areas that are smooth and relatively devoid of craters. These *mare* (Latin for seas, the seventeenth and eighteenth century interpretion from telescopic views) show ample evidence of craters that have been partly covered or obliterated by the material that makes up the smooth, dark surfaces (Figure 7.4b). The simplest interpretation is that the mare are lowland basins that were flooded by lavas sometime after an early, heavy bombardment of the Moon occurred. The flooding obliterated most of the craters, leaving a fresh surface on which some remains of old craters can be seen, and a few small, new craters were formed by impacts after the lava solidified.

Explorations by the Apollo astronauts from 1969 to 1972 returned nearly 400 kg (900 lbs) of moon rocks from mare and highland regions. Radioisotopic techniques, described in Chapter 5, were used to provide absolute dates for the solidification of these rocks from original molten materials. The lunar highlands are old, with rocks dating as old as 4.5 billion years. The mare deposits typically are 4.2 billion to 4.3 billion years old, significantly younger than the highlands.

The age estimates based on the cratering density on the lunar surface are confirmed by the absolute dating of mare and highlands provided by rock samples. It then would appear possible to use crater densities on other worlds, not accessible for sample collection at present, to construct chronologies as well. The most straightforward chronology involves determining relative ages of events on a surface, that is, which event preceded another. This simply requires counting craters as well as looking for evidence of craters partly obliterated by geologic processes. More difficult is to try to assign actual dates, which requires assuming that the lunar crater density and ages based on Moon rocks can be transferred directly to other bodies in the solar system.

7.2.1 Relative ages of events on a planetary surface

Impact craters can be used to determine the relative ages of geologic features on a planetary surface. Two examples of this are shown, one from Mars and one from Jupiter's moon Ganymede, using images from *Viking* and *Galileo* missions in Figures 7.5a and 7.5b, respectively. In the case of Mars, the geologic features of interest are channels that clearly were cut by water, but today are dry along with the rest of the planet. Are the features young or ancient? Was the climate wet up through recent times, such that life might have evolved to an advanced stage?

Examination of the *Viking* images such as that in figure 7.5a reveals that the Martian channels typically are overlain by impact craters, some fairly substantial in size. Other regions of the Martian surface have far fewer craters, and hence we can say that the channels are, relatively speaking, ancient. Determining a more exact age requires tying the cratering rate to some absolute timescale. At the same time, we know that the channels are not among the oldest features, either, because many cut through craters that must therefore be older than the channels. A chronology can be assembled in which channel formation occurs after formation of the oldest Martian terrains but before a number of other geologic events that are recorded in the surface.

Ganymede, the third of Jupiter's four giant moons (Io is closest to Jupiter, and then Europa, Ganymede, and Callisto), shows lines in its spectrum typical of water ice. However, the mass of the planet is too heavy given its volume (mass over volume is density) to be pure water ice. The best guess, based on models of solar system formation, is that the heavier component is a *silicate*, or common rocky material containing silicon, oxygen,

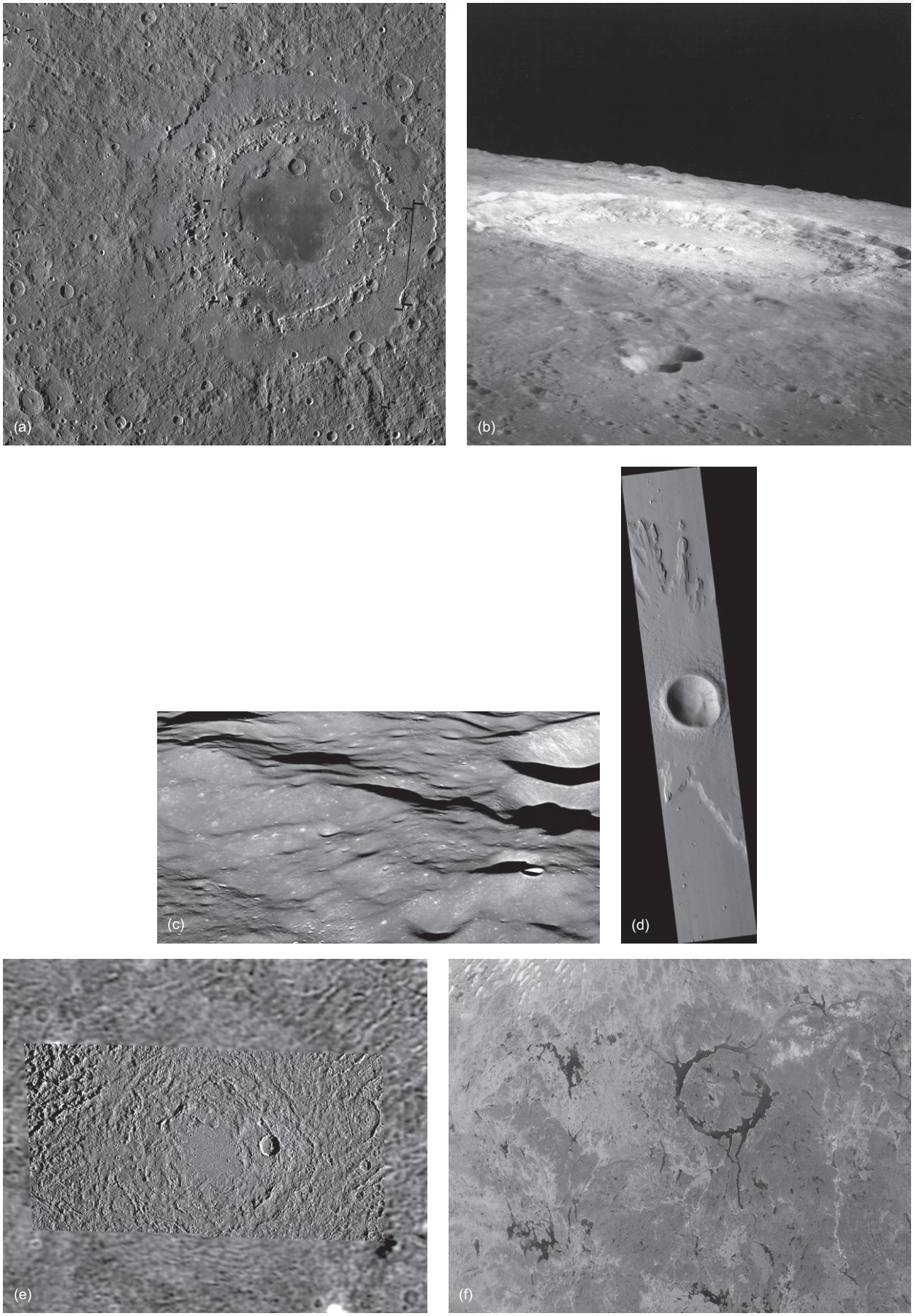


Figure 7.3 Varieties of impact craters: (a) large multiring basin, Mare Oriental, on the Moon; (b) classic large crater, Copernicus, with central peak, on the Moon; (c) smaller lunar craters without peaks; (d) pedestal crater on Mars, formed by melting of ground ice during impact. (e) relaxed craters, or *palimpsests* on Ganymede (*Voyager* image); (f) eroded crater on Earth, now comprising Lake Manicouagan, Quebec, Canada (see color version in plates section). Photos (a) through (f) are courtesy of NASA.

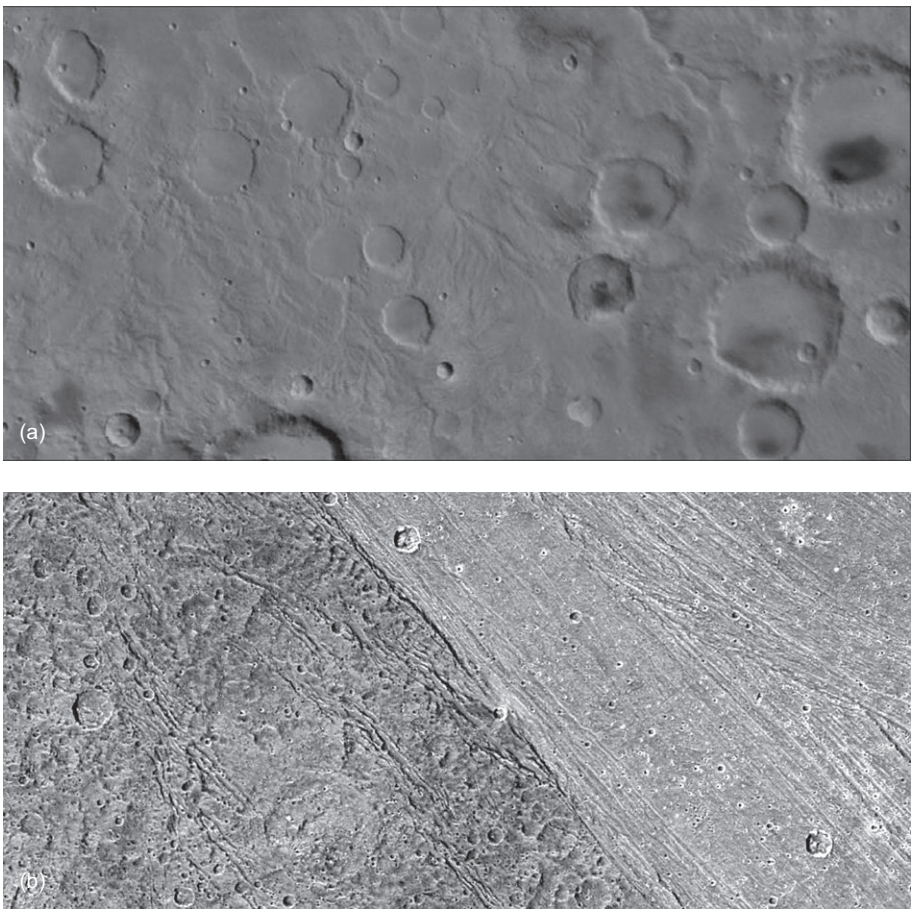


Figure 7.4 Two very different terrains on the Moon: (a) the lunar highlands show craters of all sizes filling all available surface space; (b) the lunar mare regions are smooth, dark plains with a few fresh craters and remains of large craters, in various states of preservation, which were present when lavas flooded the lunar surface.

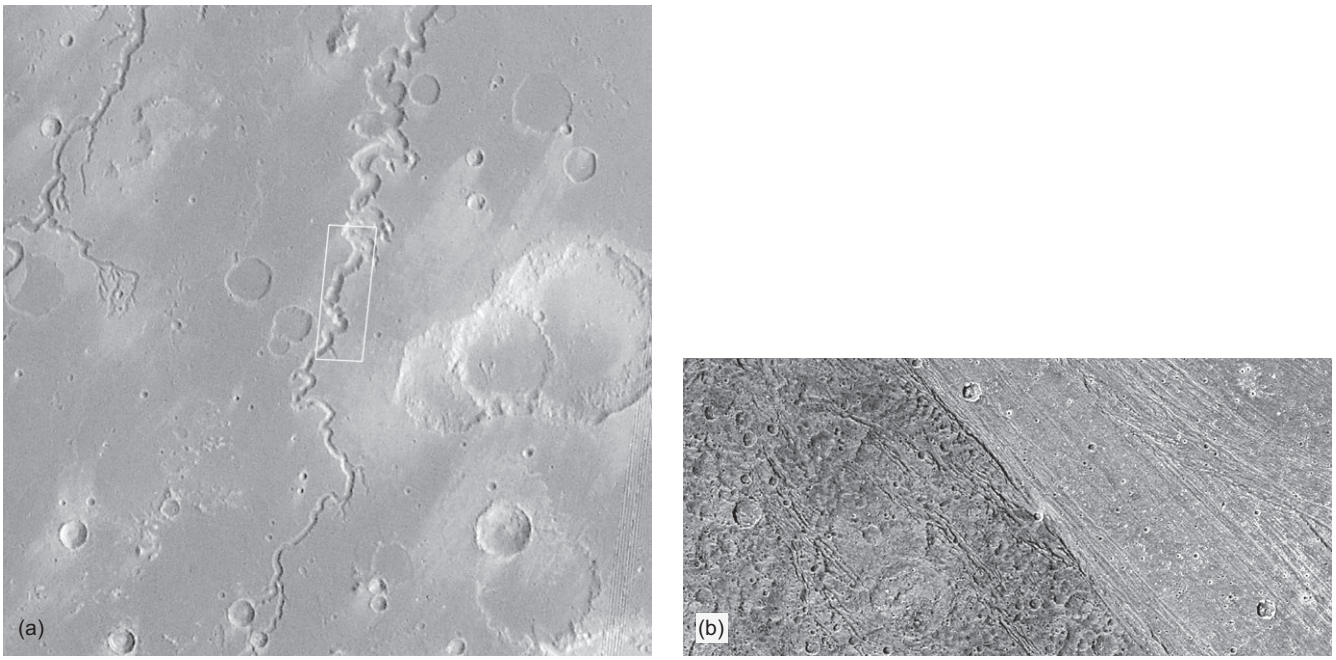


Figure 7.5 (a) NASA/*Viking* image of Martian surface cut by channels; (b) NASA/*Galileo* image of dark and light terrains on Jupiter's moon, Ganymede.

magnesium, some iron, and other elements. Therefore, unlike our own Moon, Earth, Venus, Mercury, and Mars, which are made up mostly of silicon-bearing rock and metal, Ganymede is half-rock, half-ice. This is true for Callisto, as well, but not Europa and Io: they are both mostly rock, although Europa has an outer veneer of ice and possibly liquid water.

One might expect, from terrestrial experience, that the ice might behave differently in an impact than rock. In fact, the pressures and temperatures in the hypervelocity impacts we have been describing are so large that there is little difference. Furthermore, temperatures in the distant outer solar system, where Jupiter and its moons reside, are so low that ice behaves much like rock as a material making up the solid *crusts*, or outer layers, of Ganymede and Callisto. The surface temperature near the equator of these moons is typically 165 K, very far below the ice melting point of 273 K.

Images of Ganymede reveal two types of surfaces: a dark, heavily cratered terrain, and a bright, lightly cratered terrain. The paucity of craters on the latter surface immediately suggests that it is a younger feature, perhaps ice that has been extruded from the interior along cracks and flowed outward. In places, it is possible to see where a crater on the older dark terrain has been partially obliterated by the new material. It is also possible to tell something about how the cracks and new material formed by looking at distortions in partially preserved craters along the edges of the bright terrain. The difference in brightness between the dark and light terrains remains a matter of speculation; silicates and perhaps some carbon-bearing materials are well mixed with the water ice, perhaps dating back to the original formation of Ganymede.

The ability to learn something about the sequence of events on a surface by looking at crater densities is a tool of primary importance in solar system studies. It is a new development of a much older technique applied to Earth geology to look at *superposition* of layers to assemble a history of a given region. On Earth, water and geologic activity have effectively erased the cratering record, so that the use of craters as a geologic tool was a novel idea that did not come into its own until planetary exploration began some three decades ago.

7.2.2 Absolute chronology of solar system events

Relative age dating is limited in the amount of information derived. Ideally, one wants to assign ages to events on the surfaces of planets and moons so as to understand their history and ultimately that of the solar system. Imagine how limited our own understanding of the history of human cultures would be if we only knew the order of events, but not their antiquity or duration.

In the case of Earth's *geologic* history, even before radioisotopic dating provided reliable dates, estimates of ages could be made on the basis of notions of the accumulation rate of *sediments*, debris brought from high to low places by the action of water. Early work tended to overestimate the rates of sedimentation and hence produced a compressed timescale relative to what is accepted today based on radioisotopic determinations. With the help of radioisotopic dating, the rates of geologic processes are now better understood and calibrated, such that indirect dating techniques such as sedimentation are enhanced as tools in assembling the history of Earth.

The situation for an absolute chronology from planetary cratering is similar in that radioisotopic dating has been used to construct a chronology for Earth's Moon, which then has been applied, with caveats, to other solar system bodies. The Moon is the only body for which radioisotopic dating of terrains of varying crater density can be performed; Earth's cratering record is too sparse. (It is not possible to determine unequivocally from what part of the Martian surface the SNC meteorites were derived; hence they cannot help calibrate the cratering record on Mars.)

The oldest parts of the Moon, the highlands, have by far the largest number of craters; the younger mare possess the least. This is consistent with the decreasing population over time of debris in orbits around the Sun. Theories of planet formation, which we discuss in Chapter 10, hold that the planets were assembled from smaller pieces of rock and ice through relatively low-speed collisions that allowed the pieces to stick together. In the final phases of this process, most of this *protoplanetary* material was perturbed by close encounters with the planets into highly elliptical orbits, guaranteeing that any subsequent collisions with the planets would be at high speeds, producing craters. Over time this remnant debris of planet formation was swept up by the planets, so that the available impactor population has decreased dramatically from the beginning to the present day.

A simple law governing the rate of impacts over time, consistent with the sweep-up picture described above, and with the lunar cratering record, has the *inverse exponential* form shown in Figure 7.6. The curve is characterized by a very steep decrease initially, as large amounts of material are swept up by the nearly fully grown planets, followed by a transition to a slowly decreasing rate of impacts. The cratering record on the Moon tells us when the transition occurs between these two regimes. Further, it provides information about the tail-off in impacts at later times, though with limited capability because of the paucity of new craters. More difficult to discern is the precise steepness of the early curve, because the cratering rate was so high that lunar highland surfaces are completely covered with craters: new impacts simply obliterate all or part of old ones and only a lower limit on the ancient cratering rate is accessible.

The dating of Moon rocks fixes the transition in the cratering curve at roughly 3.8 billion to 4.0 billion years before present; the period of intense cratering before that is called the *Late Heavy Bombardment*, referring to the tail end of the planet-formation (*accretion*) process. Interestingly, the oldest whole rock samples on Earth date back to roughly the same time. We know that this does not represent the age of Earth because the rocks are rather evolved, showing the action of liquid water on their chemistry and texture; additionally, meteorites record much earlier dates back to 4.56 billion years before present. Instead, Earth was simply too active geologically at earlier times to preserve older rocks and, as we see in later chapters, had little or no continental land mass on which such rocks could be preserved.

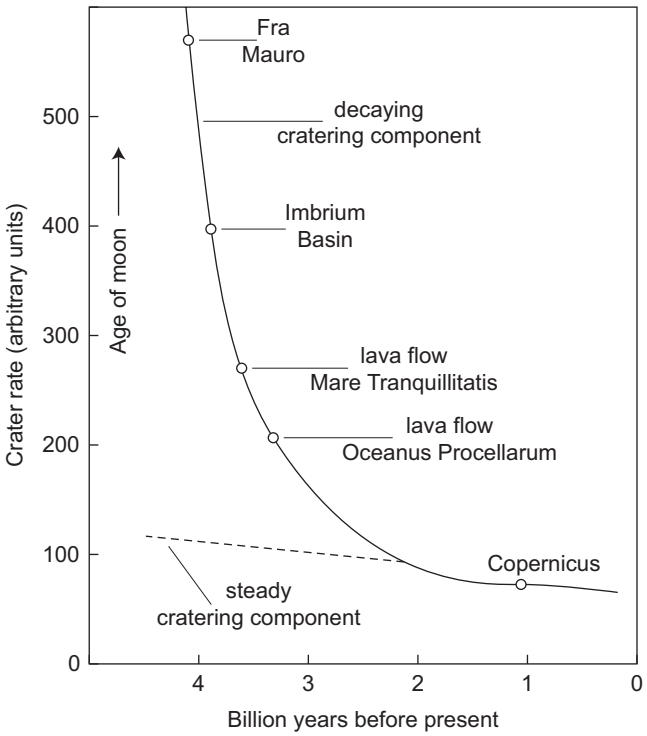


Figure 7.6 Number of impacts versus time on the surface of the Moon; the curve is labeled with ages of rocks collected at the *Apollo* landing sites, and an estimate for the age of the large crater Copernicus. The dashed line shows what the cratering rate would look like if the recent cratering rate were extrapolated along a straight line back to the beginning. After Lang and Whitney (1991).

An additional piece of information on the bombardment history of the Moon, one crucial for calibrating the impact history of other solar system bodies, is the distribution of crater sizes. Crater sizes are related fairly directly to those of the original impactors, and hence a model of crater formation can yield the original impactor size distribution.

A typical crater population will exhibit the evolution sketched in Figure 7.7. On a plot showing the number of craters below a given size, versus size, the data tend to fall on a broken line with two different slopes. The steeper slope, occurring for larger crater diameters, directly reflects the size distribution of the original impacting population that produced the craters. At smaller crater sizes, *saturation* effects tend to reduce the number of observed craters: Smaller craters are so numerous that they readily fill a surface until new craters simply obliterate the old ones. Over time, as shown in the figure, the breakpoint at which saturation takes over moves to larger crater sizes as the surface is increasingly filled. Determination of the breakpoint on such a plot for a cratered region of a planetary surface provides a measure of its age when correlated against surfaces that are absolutely dated, such as those of the Moon.

Mercury and Mars show heavily cratered terrains with distributions similar to those on the Moon. Not only does this allow us to determine how ancient the various terrains are, it also leads us to conclude that the impactor populations on the Moon, Mars, and Mercury are similar. This strongly suggests that the population of impactors that have struck the Moon over time originate from beyond Earth orbit, and in fact are in orbits around the Sun that take them well beyond Mars into the outer solar system. Some of these impactors may have been icy bodies derived from reservoirs of debris beyond the orbit of Jupiter, and left over from planetary formation. Additional impact debris likely is derived from the asteroid belt between Mars and Jupiter.

The crater distributions on the moons of the giant planets are similar neither to those of the Moon, Mars, and Mercury nor to each other. Each giant planet seems to have defined a unique population of impactors for its moons, with the only general resemblance that the population has decreased sharply with time. The cratering size-frequency distribution for the moons of

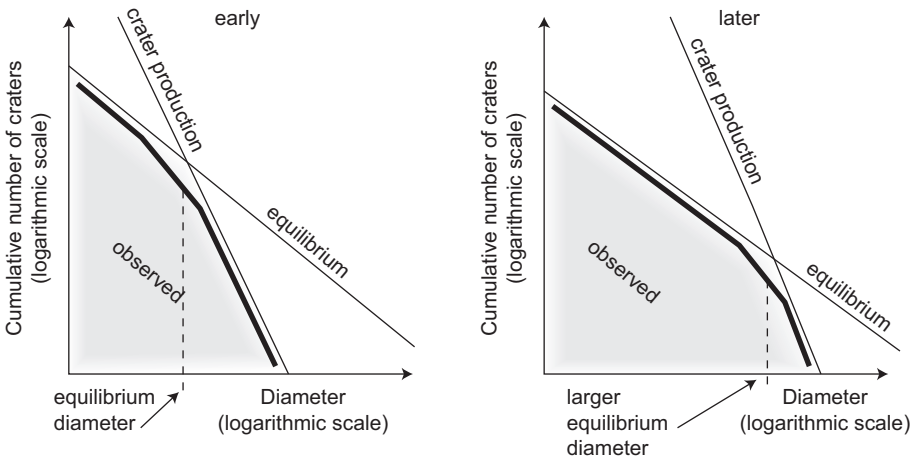


Figure 7.7 How a population of craters evolves over time. Shown in each graph is the number of craters with a diameter less than D, as a function of D. The thinner lines are guides to two idealized populations of craters. The steeper line, "production," is what is produced directly from a particular size distribution of the impacting bodies. The shallower "equilibrium" line is the result of saturation, i.e., obliteration of craters by newer impacts in a very crowded crater field. The right-hand panel represents the situation at a time later than the left-hand panel, showing that, as a surface gets older, the effects of saturation extend to larger and larger crater diameters. Redrawn from Melosh (1989, p. 192).

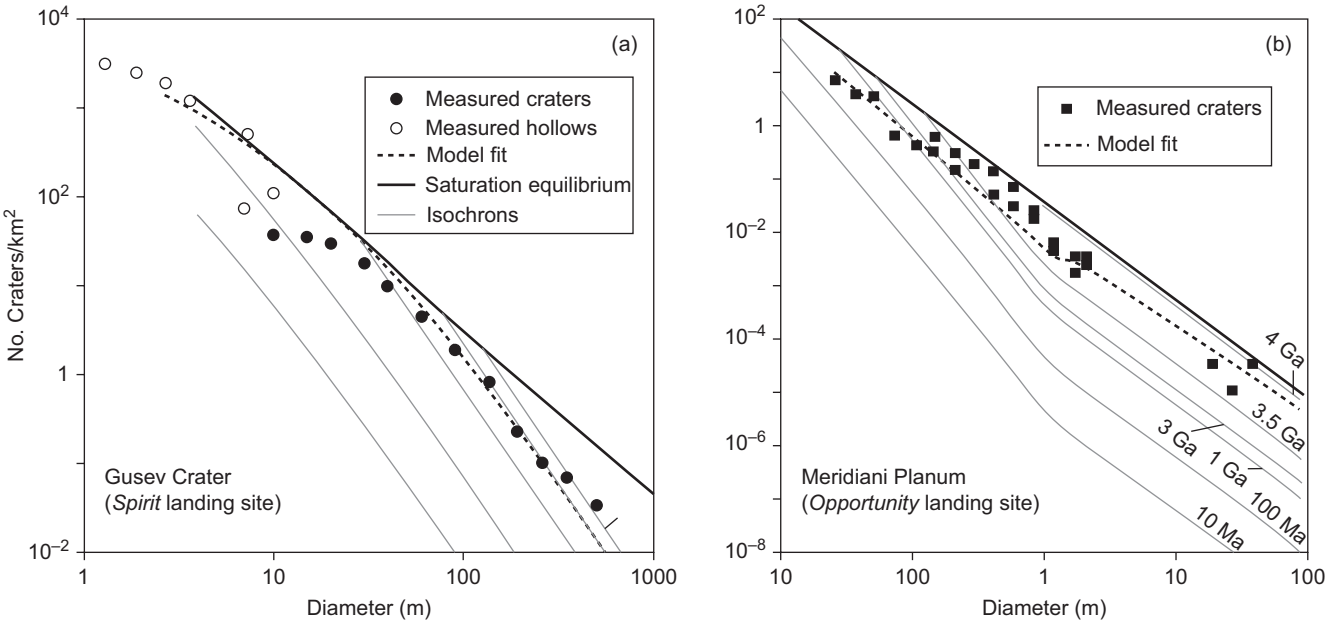


Figure 7.8 Examples of crater size-frequency distributions for the two sites on Mars where the Mars Exploration Rovers ranged over the surface. Plotted are the cumulative number of craters per square kilometer versus diameter. (a) At the *Spirit* site, a large number of hollows exist which may be eroded or partly buried craters; these are indicated with empty circles. The saturation curve defined in Figure 7.7 is shown as a heavy black line, and model fits for different ages in millions of years (Ma) and billions of years (Ga) are shown as lines of constant age ("isochrones"). At the *Spirit* site a classic fit for an ancient terrain, in which the smaller craters are saturated while the larger ones provide an age, is obtained (dashed line). (b) For the *Opportunity* site, which is suspected to have been more eroded by the action of water, the fit is poorer, presumably because the crater are not uniformly preserved and many are unrecognizable or obliterated. From Smith *et al*. (2008).

the giant planets probably can be understood best by invoking two populations of impactors: those in solar orbit, perhaps the same as those that had peppered the inner solar system with craters, and a unique population of debris orbiting each of the giant planets. This local debris, likely the leftovers from the formation of the giant planets and their moons, has had a different history for each of Jupiter, Saturn, Uranus, and Neptune. In the case of Saturn, there is even evidence in the crater record for the break up of a large moon late in solar system history, perhaps in the orbit now occupied by the irregularly shaped moon Hyperion.

The cratering record in the outer solar system is a useful tool for assembling a rough history of this region but, as yet, it is too remote to retrieve samples of icy moons for radioisotopic dating. From studies of craters we now know that Callisto's heavily cratered surface probably dates back close to the beginning of the solar system; the heavily cratered terrain on Ganymede might be slightly younger but certainly predates the bright lightly cratered regions on that moon. Saturn's small moon, Enceladus, exhibits smooth regions bordered by areas where craters have been partly covered by bright flows; clearly this moon has been active enough that fresh water ice and other compounds poured out onto the surface recently. Uranus' satellites differ in their crater density and hence bespeak varying levels of geologic activity the nature of which is otherwise poorly understood. Finally, Jupiter's Io is devoid of craters and the same is nearly true of Jupiter's Europa and Neptune's Triton, satellites that possess a history whose levels of geologic activity rival or exceed that of our Earth in erasing the cratering record.

7.3 Cratering on planetary bodies with atmospheres

In our discussion of the inner solar system, the reader might have noticed that we omitted mention of Venus. That planet's surface lies beneath an enormous atmosphere with a surface pressure 90 times that of Earth. The geology was finally mapped with completeness by a radar imager aboard the US *Magellan* spacecraft beginning in 1989. Impact craters are sparse, indicating a relatively youthful surface renewed by volcanism, likely similar to the basaltic types of volcanism on the Earth, a near twin in terms of size and mass. However, most striking is the predominance of large impact craters relative to the distributions seen on airless bodies.

The thick atmosphere of Venus acts as a shield, preventing smaller impactors from reaching the surface intact. The pressure of the atmosphere, close to the surface where it is densest, exerts a stress on the incoming material. Most comets and even asteroids are relatively weak; that is, their interiors have been fractured and hence are not strongly glued together. The effect of aerodynamic forces is to crush the impactor; the individual fragments then spread out. Very small impactors may spread out well above the surface, the individual fragments burning

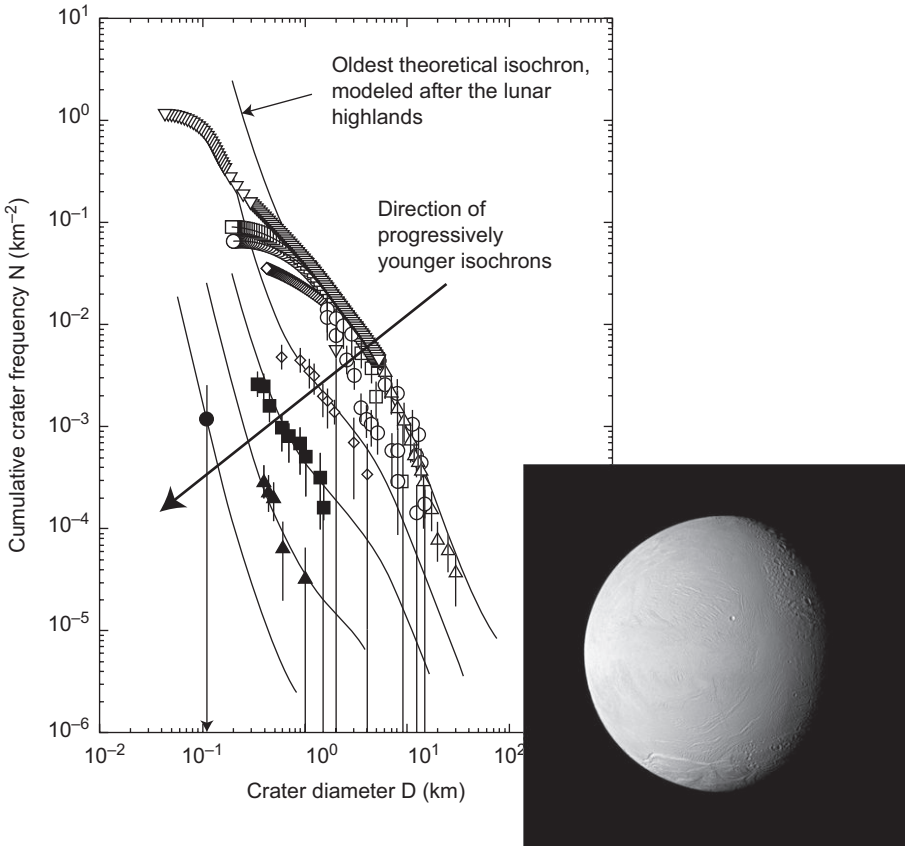


Figure 7.9 Cumulative crater frequency versus crater diameter in km on Enceladus. Because Enceladus exhibits terrains with widely different crater densities (inset), these regions have been plotted separately from each other. They each follow a different isochron, which have been drawn in without ages because there is no absolute calibration for the outer solar system in the same way that there is for the inner solar system (where lunar rock samples exist). The lunar highlands, which are nearly as old as the Moon itself (Chapter 5), have been used as the model for the isochron representing the oldest portion of Enceladus. The isochron has been scaled in absolute position on the graph based on estimates for the relative cratering rates in the inner versus the outer solar system; evidently even the oldest terrain has undergone some modification and removal of the smallest craters. Modified from Porco *et al*. (2006), with inset courtesy NASA/JPL/Space Science Institute.

up in the atmosphere; this is the likely fate of the small asteroid that exploded above the Tunguska River in Siberia in 1908, knocking down people 60 km from the impact site but leaving no crater. Very large bolides are hardly affected by the atmosphere and leave a normal crater. Intermediate-size impactors disrupt and partially disperse, leaving an oddly shaped crater or a field of smaller craters clustered close together.

The size thresholds defining the different results of an impact into an atmosphere depend to some extent on the strength of the impactor, but much more importantly on the thickness of the planet's atmosphere. This leads to the intriguing possibility that one could use the crater distribution on ancient terrains to learn what the atmosphere of a planet was like a long time ago. If Venus, for example, had a much thinner atmosphere in its past, ancient terrains should show a crater size distribution akin to that of the Moon. Likewise, Mars may have had a thick atmosphere in the past, and perhaps the most ancient terrains record evidence for this (Figure 7.8). Unfortunately in the case of Venus, the surface has been so geologically active that no ancient terrains appear to be present, making it impossible to use the crater record as a probe of the early atmosphere.

It may be possible to apply this approach to Saturn's moon, Titan, which is larger than the planet Mercury and has an atmosphere at its surface four times denser than the Earth's air at sea level. In consequence, Titan's atmosphere is intermediate in thickness between Earth's and that of Venus, and filters out small impactors rather effectively. Some models of the evolution of this moon suggest that its atmosphere might have been much thinner several billion years ago when the Sun was less luminous (see Chapter 14); if this turns out to be the case, craters too small to have been produced during the current thick-atmosphere epoch should be present. Our first detailed glimpse of the surface, hidden behind a global haze, came from the *Cassini–Huygens* mission, which arrived in 2004 and has mapped about 30% of the surface of this giant moon. The cratering record suggests an age of about 1 billion years for Titan's surface, but the presence of an atmosphere and blankets of organic debris make a more precise determination challenging. For the other Saturnian moons, the absence of substantial atmospheres provides a better opportunity to map surface ages (Figure 7.9).

7.4 Impactors through time

In our discussion of the dating of events through the cratering record, we noted that underlying the impactor flux is the assumption that collisions are winding down as the solar system is slowly cleared of debris. Although this general concept is a useful one, it is important to consider that sources of collisional debris for the inner solar system continue to be supplied over time as comets are perturbed out of their orbits in the Kuiper Belt and Oort Cloud and travel inward to possible collisions with the planets and their moons.

It has been suggested that, occasionally, large numbers of such comets may be perturbed into the realm of the terrestrial planets, temporarily increasing the impact rate and producing comet showers. Although the evidence for such major events is tenuous (and the idea comes in and out of vogue), the more general notion that the flux of comets through the inner solar system is variable is sensible. The consequences of a collision of a comet-sized body with Earth are profound for the geology, climate, and biology of this planet. We explore these consequences in Chapter 18.

Summary

Only one place in the solar system, the Moon, has been sampled to the extent necessary to create an absolute chronology of its history. For the rest of the solar system, relative age dating must be used to construct the sequence of events that have occurred as exhibited by surface features. Impact craters are the most ubiquitous and readily recognizable features to use for such relative dating. Not all craters are formed by the hypersonic impact of projectiles into planetary surfaces; collapse pits and volcanic explosion craters are other examples. The process of impact crater formation involves the deposition of mass and kinetic energy from the projectile into the target; for large impactors the target behaves like a fluid, generating waves in the "solid" material and endowing the impact site with not only a crater but also a central peak and possibly multiple extended rings as well. Over time craters are eroded by various geologic processes, including wind and water erosion, burial by volcanic flows, and even space weathering, smoothing the surface and rendering it more "youthful". The density of craters, that is, number per unit area, is thus a measure of the age of the surface. The superposition of craters atop other features such as fractures, or the cutting of craters by such features, allows the sequence of events to be determined. If the overall crater density of the surrounding surface can be determined it is possible to tell whether a geologic event was recent or ancient. Only in the case of the Moon is an absolute chronology available, but the lunar chronology can be extrapolated through the inner solar system with reasonable precision to determine the approximate absolute age of surfaces or timing of events. For the outer solar system, this is much more difficult, because the impactor population may have been different in the outer solar system than at the Moon.

Questions

1. Could the cratering record on Titan be used to determine whether that moon of Saturn ever had an atmosphere much thicker than it does today? If so, how?
2. Why is it necessary to use lunar samples to determine the cratering rate as a function of time in the inner solar system? Is there a reliable way to do so without physical samples?
3. When one says that a surface is youthful because it lacks craters, does it matter whether the mechanism of crater removal is volcanic versus burial of craters by sediments? Do the two mechanisms imply something different about the history of any particular planet?
4. How might one determine, from the cratering record of an icy satellite, whether most of the impactors came from solar orbit (around the Sun) or had been in orbit around the moon's parent planet?

References

Bloom, A. L. 1978. *Geomorphology: A Systematic Analysis of Late Cenozoic Landforms*. Prentice-Hall, Englewood Cliffs, NJ.

Engel, S., Lunine, J. I., and Hartmann, W. 1995. Cratering on Titan and implications for Titan's atmospheric history. *Planet Space Science* **43**, 1059–66.

Lang, K. R., and Whitney, C. A. 1991. *Wanderers in Space: Exploration and Discovery in the Solar System*. Cambridge University Press, Cambridge, UK.

Melosh, H. J. 1989. *Impact Cratering: A Geologic Process*. Oxford University Press, New York.

Porco, C. C. *et al*. 2006. Cassini observes the active south pole of Enceladus. *Science* **311**, 1393–1401.

Smith, M. R., Gillespie, A. R., and Montgomery, D. R. 2008. Effect of obliteration on crater-count chronologies for Martian surfaces. *Geophysical Research Letters* **35**, doi:10.1029/2008GL033538.

Spudis, P. 1993. Moon, geology. In *Van Nostrand's Scientific Encyclopedia* (D. M. Considine, ed.). Van Nostrand Reinhold, New York, pp. 452–5.

8

Relative age dating of terrestrial events: geologic layering and geologic time

Introduction

Prior to the invention of radioisotopic techniques for dating rock samples, geologists determined relative ages for rocks using simple principles of how rocks and their fragments are deposited, and using remains or records of extinct life to correlate samples from different locations. When combined later with the dating of rocks by radioisotopic techniques, a detailed history of Earth could be developed. We work with this history repeatedly throughout the rest of the book. This chapter serves as an introduction to the techniques used to assemble such a record.

8.1 Catastrophism versus uniformitarianism

When we look at Earth's landforms, we are viewing a snapshot, a moment in a vast span of time during which mountains rise and fall, seas expand over land areas and contract again, and continents shift their positions and grow slowly from new rock added by volcanoes. These processes all require vast amounts of time for their completion, but most do not proceed in a smooth, gradual manner. Instead, geologic processes are a combination of gradual effects and sudden catastrophes. The earthquakes that shake California represent sudden failures of rock after the build up of stresses over time as one portion of California slowly glides past the other, as we discuss in Chapter 9.

The realization that Earth changes in this hybrid fashion was long in coming. Much of the history of the development of geology was a battle between those who argued in favor of *uniformitarianism*, and hence gradual change over enormous spans of time, and those who claimed that Earth was young and shaped by catastrophic processes. As estimates for the age of Earth climbed, it appeared the uniformitarians were right. In the past few decades, though, the importance of catastrophic change on Earth has become clear, in large measure through study of other planets. However, radioisotopic dating (Chapter 5) confirmed a large age for Earth, and so, both camps were right.

The long history of this debate has a literature all of its own, and it is appropriate only to touch on some aspects here, to illustrate the ways people attempted to gauge Earth's age, and as an introduction to geologic processes.

8.2 Estimating the age of Earth, without radioisotopes

One of the first recorded observers to surmise a long age for Earth was Herodotus, who lived from approximately 480 to 425 BC. He is best known as the father of history, having written an extensive account of the Persian invasion of Greece culminating around the time of his birth. Herodotus was also a traveler, and visited the Nile River Valley, which was subject to annual cycles of flooding. He came to the important conclusion that the Nile Delta was in fact a series of sediments built up in successive floods. By noting that individual floods deposit only thin layers of sediments, he was able to conclude that the Nile Delta had taken many thousands of years to build up. (In fact, Herodotus coined the term "delta" for the accumulation of sediments at the mouth of a river; the shape of the Nile Delta reminded him of the Greek letter by that name, Δ.)

More important than the amount of time Herodotus computed, which turns out to be trivial compared to the age of Earth, was the notion that one could estimate ages of geologic features by determining rates of the processes responsible for such features, and then assuming the rates to be roughly constant over time. Similar applications of the principle of uniformitarianism were to be used again and again in later centuries to estimate the ages of rock formations, and in particular of layers of sediments that had compacted and cemented to form *sedimentary rocks*.

Throughout the Middle Ages, European studies of the history of Earth relied on the Bible, and it wasn't until the seventeenth century that attempts were made again to understand clues to Earth's history through the rock record. Nicolaus Seno

(1638–1686) worked out principles of the progressive laying down of sediments in Tuscany. However, a Scottish doctor and farmer, James Hutton (1729–1797), was the first to have the important insight that geologic processes are cyclic in nature. Forces associated with subterranean heat, which we deal with in Chapter 9, cause land to be uplifted into plateaus and mountain ranges. The effects of wind and water then break down the masses of uplifted rock, producing sediments that are transported by water downward to ultimately form layers in lakes, seashores, or even oceans. Over time, the layers *lithify* to become sedimentary rock. These are then uplifted sometime in the future to form new mountain ranges, which exhibit the sedimentary layers (and the remains of life within those layers) of the earlier episodes of erosion and deposition.

Hutton's concept represented a remarkable insight, because it unified many individual phenomena and observations into a conceptual picture of Earth's history. With the further assumption that these geologic processes were generally no more or less vigorous than they are today, Hutton's examination of sedimentary layers led him to realize that Earth's history must be enormous, that geologic time is an abyss and human history a speck by comparison.

Particularly inspiring to Hutton was Siccar Point, in Scotland, where the sedimentary record was interrupted by an *unconformity*, in which a sequence of layers is missing because of erosion and removal; the layers above and below usually are tilted relative to each other by the tectonic forces that produce uplift and build mountains. Hutton wrote:

> Here are three distinct successive periods of existence, and each of these is, in our measurement of time, a thing of indefinite duration . . . The result, therefore, of this physical inquiry is, that we find no vestige of a beginning, no prospect of an end.

After Hutton, geologists tried to determine rates of sedimentation so as to estimate the age of Earth from the total length of the sedimentary, or *stratigraphic*, record. Typical numbers produced at the turn of the twentieth century were 100 million to 400 million years. These underestimated the actual age by factors of 10 to 50 because much of the sedimentary record is missing in various locations (due to erosion) or compressed (due to metamorphism; see section 8.4) and because there is a long rock sequence prior to half a billion years ago that is far less well defined in terms of fossils (see section 8.6) and less well preserved.

Various other techniques to estimate Earth's age fell short, and particularly noteworthy in this regard were flawed determinations of the Sun's age. It had been recognized by Immanuel Kant (1724–1804) that chemical reactions could not supply the tremendous amount of energy flowing from the Sun for more than about a millennium. Two distinguished physicists, Herman von Helmholtz in the mid-nineteenth century, and William Thompson (Lord Kelvin) at the end of the nineteenth century both came up with lifetimes based on the Sun's energy coming from *gravitational contraction*. Under the force of gravity, the compression resulting from a collapse of the object (whether slowly or quickly) must release potential energy (Chapter 3). Ages in the tens-of-millions-of-years range were derived, much less than the geologic estimates of the time.

Although, at present, the Sun's energy is not derived from gravitational contraction, this is a primary source of energy during star formation, when interstellar gas and dust collapse from a more diffuse state (Chapter 10), and during some stages of the end of a star's life. This holds true for the planets as well, whose late stages of formation were characterized by material falling at high speeds under the influence of the gravitational fields of the growing planets. For Earth, temperatures during this *accretion* process were likely sufficient to melt rock. Some of the heat that we measure today coming from interiors of Earth and other planets is in fact a remnant of this initial energy of collapse.

It was the discovery of radioactivity at the end of the nineteenth century, independently by French scientists Henri Becquerel and Marie Curie and German physicist Wilhelm Röntgen, that opened the door to a solution of both the Sun's energy source and the age of Earth. From the initial work came a suite of discoveries leading to radioisotopic dating, which quickly led to the realization that Earth must be billions of years old, and to the discovery of nuclear fusion as an energy source capable of sustaining the Sun's luminosity for that amount of time. By the 1960s, analysis of meteorites and refinements of solar evolution models both converged on an age for the solar system, and hence for the Earth, of 4.5 billion years (Chapter 5). Figure 8.1 summarizes the steps in the increasing estimates of the age of Earth.

8.3 Geologic processes and their cyclical nature

Rocks that comprise Earth's surface originate in the interior as molten or semimolten material (Figure 8.2). Some rocks find their way to the surface in volcanic eruptions, in which case the solidified rock is a *volcanic igneous* rock. Most continental igneous rocks follow a different route: they may cool and solidify beneath the surface, or in the hidden cores of mountain ranges, as *plutonic igneous* rocks. These may be exposed eventually as elevated terrains such as plateaus or mountains are eroded by wind and water. Often, such igneous rocks will cut through sedimentary or metamorphic (defined later in this section) layers in the final stages of their elevation and solidification. Note that oceanic rocks originate almost entirely from undersea volcanism and have a distinctly different composition from continental rocks; we discuss these differences and their origin in Chapters 11 and 16.

Once rocks are exposed at the surface they are subjected to *weathering* processes, which include abrasion and erosion by wind, and, much more important, erosion by water. Water molecules in raindrops will combine chemically with atoms comprising the minerals in the rock, weakening the material and accelerating its crumbling. Water works its way into cracks in the rock and expands when it freezes; this leads to mechanical breaking of rock. Liquid water carves out river valleys, often along pre-existing fractures or *faults* in the rock. Ice sheets, or glaciers, can do the same in colder regions. Some minerals such as limestone are dissolved directly by the action of water, forming caves and exotic-looking structures known as *Karst* topography. The removal of material eroded from the uplands is

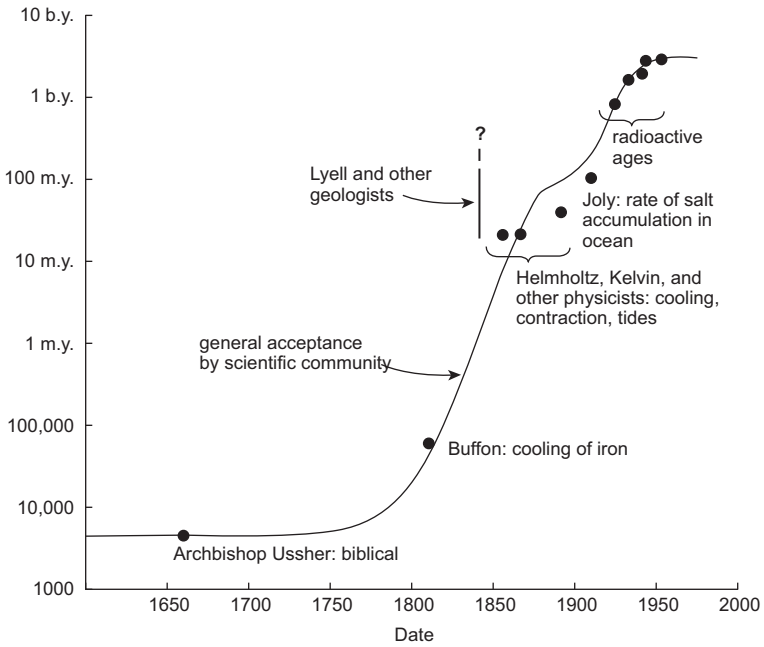


Figure 8.1 Age of Earth as estimated by various techniques since the Renaissance. Buffon tried to use rate of cooling of iron from a molten state to estimate the age of Earth, but the omission of as-yet undiscovered radioactive heating of Earth's interior (Chapters 9 and 11) seriously shortened his cooling time. Joly worked out how long it would take to bring the oceans up to their current salinity based on the rate at which rivers carry salt to the sea; he ignored the precipitation of salt out of the ocean water into seafloor sediments. From Press and Siever (1978) by permission of W. H. Freeman and Company.

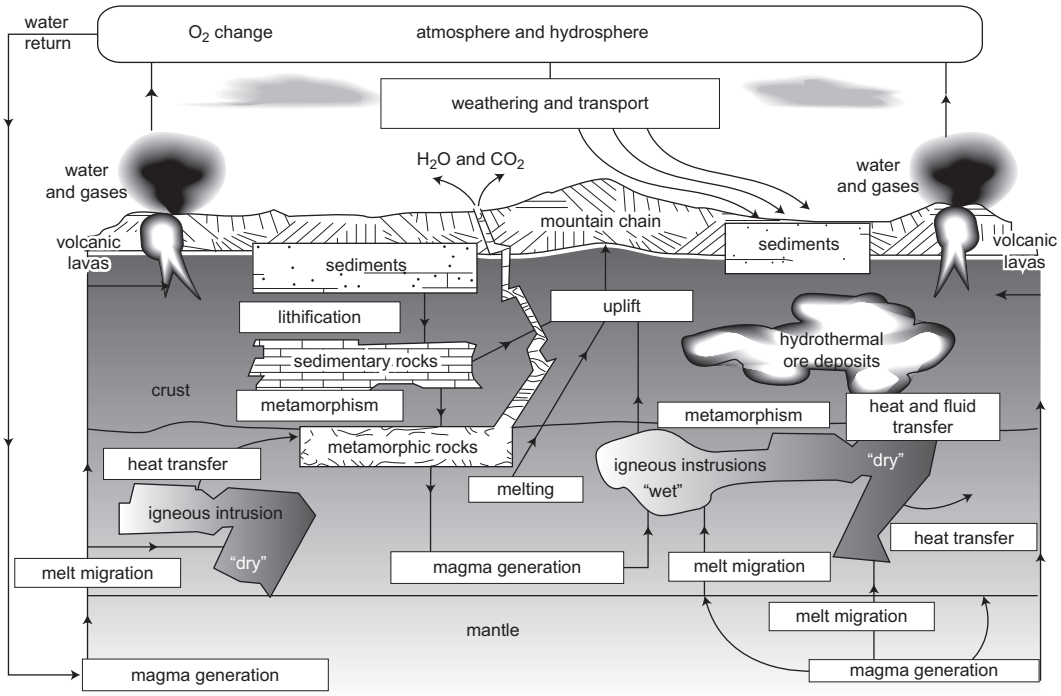


Figure 8.2 Rock cycle of Earth shown schematically. Adapted from Wyllie (1971, p. 48) by permission of John Wiley and Sons, Inc.

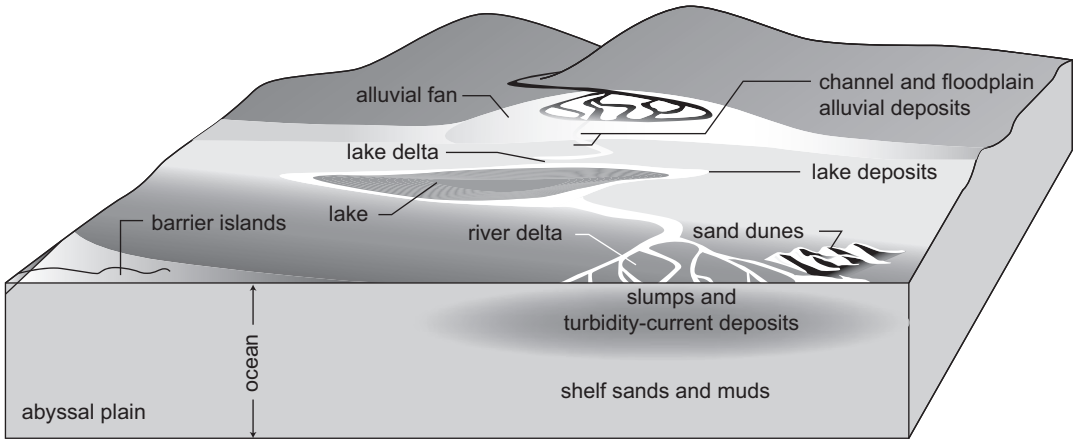


Figure 8.3 A sedimentary journey: sketch of the process of sedimentation from highlands to deposition in lakes or seas. Adapted and redrawn from Press and Siever (1978).

by streams and rivers, moving debris downward toward lakes and oceans. In regions bordering oceans and other saltwater bodies, corrosive salt injected into the atmosphere via evaporating sea spray enhances erosion.

The long journey of sedimentary debris from highlands to the sea, sketched schematically in Figure 8.3, is where geology most commonly and pervasively binds itself to the history of humankind. We rely on streams, rivers, lakes, and groundwater to sustain our existence. Agriculture began and continues to flourish in river valleys. The edge of the sea has always held a special place in human imagination both as a source of sustenance and a gateway to faraway lands.

The geologic cycle does not end with the deposition of sediments in alluvial plains or on the ocean floor. As sediments accumulate, the underlying material slowly cements to form sedimentary rocks in which the layering and pebble sizes and shapes are preserved. These rocks may be exposed by uplift combined with erosion and usually are tilted in the process. Erosion will remove some layers, but eventually, the scene changes enough that new sediments are deposited atop the old; missing sedimentary layers removed by erosion as well as the tilt of the older layers disappearing suddenly at the interface with the newer deposits are easily recognizable to geologists as unconformities.

Often, sedimentary layers are so deeply buried that they are subjected to high pressures and temperatures. The sedimentary layers may be softened by the temperature, distorted by the pressures, and chemically altered to a greater or lesser degree. These rocks, uplifted once again in a cycle of mountain building, are known as *metamorphic rocks*; those only slightly altered from sedimentary rocks are known as *low-grade* metamorphics, whereas those heavily altered are *high grade*.

Some sedimentary layers, particularly on the seafloor, eventually return to the interior of Earth. These rocks are melted, mixed with other molten and solid rocky material, and ascend as new lavas (via volcanoes) or magmas (via plutonic igneous processes) to start the cycle over. The new igneous rocks, having been cycled through the crust and surface of the earth, generally have a slightly different composition than their progenitors; this is part of the process by which continental rock with its distinct composition is created from other rock types.

8.4 Principles of geologic succession

Many of the above-mentioned geologic processes deposit layers of rock one atop the other. This is particularly true for sedimentary rocks, but one may see layering in igneous rock associated with successive episodes of volcanic eruptions. Furthermore, metamorphic rocks often are sufficiently well preserved that the original layering of the precursor sedimentary rocks is retained. Distinct rock layers are called *formations*, and sets of formations constitute *strata* or a *stratigraphic section*.

These strata are, in essence, a relative chronology of geologic events on Earth, a point that was realized by Hutton and others in the nineteenth century. It was at that time that a very basic principle, that of *geologic succession*, was developed: *the chronological sequence of any stratigraphic section depends upon the original order in which the formations are laid down.*

The application of this principle, of course, is complicated. Tectonic processes will tilt or entirely invert a set of layers. Erosion will remove whole sequences of strata, leaving an enormous gap of time between the remaining layers and those subsequently deposited above them. Igneous rock may intrude into sedimentary layers, and the layers themselves could be buried and metamorphosed.

Figure 8.4 shows an idealized example of the inference of relative ages in a stratigraphic sequence. Consideration of the geologic processes described earlier in this chapter allows one to sort out what has happened. Age decreases, generally, from the bottom to the top. The lowest layer, once a sediment, was buried and metamorphosed at some time in the past, but not enough to completely obscure the pattern of the original sedimentary layers. Also seen in the figure is an unconformity, where the lower layers were tilted, and erosion then formed a level surface on which newer sedimentary layers were deposited. Erosion could have removed a number of layers between the tilted and the level sequences, but this cannot be discerned from the figure; in section 8.6, we describe how fossils provide information on

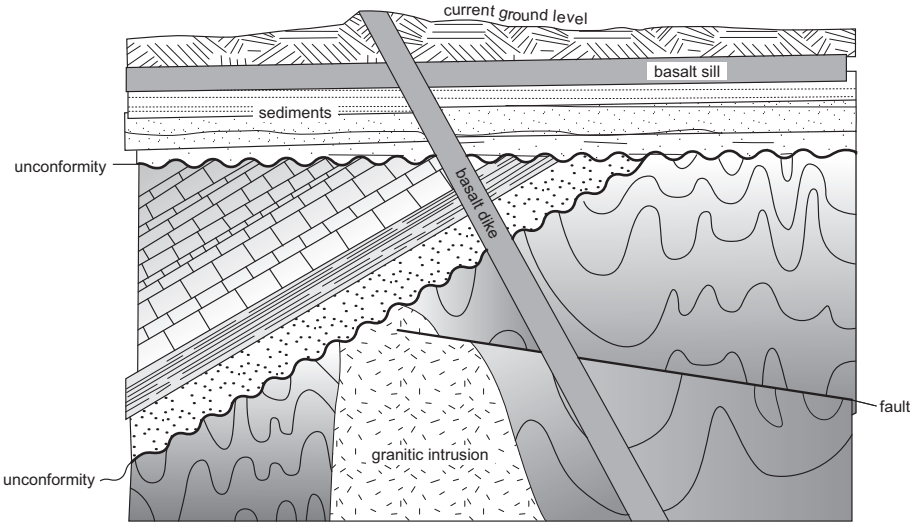


Figure 8.4 Idealized section of a geologic column from which a particular sequence of geologic processes may be inferred as described in the text. The section might be exposed, for example, as a roadcut, or as a cliff face on the side of a mountain. Adapted from Press and Siever (1978) by permission of W. H. Freeman and Company.

such gaps. Intrusions of igneous rocks along faults (*dikes*) are sketched on the figure. Clearly, they must postdate those sedimentary or metamorphic layers through which they cross, and must predate those layers that lie above them and truncate them. With these guidelines, the reader should ponder Figure 8.4 and work out the temporal sequence.

8.5 Fossils

The determination of relative time for events – that is, the order in which rock sequences were laid down or intruded into preexisting layers – provides only limited information on the history of Earth. It is difficult or impossible from the rocks themselves to reliably correlate a sedimentary layer or sequence from one part of the Earth to another, or even from one part of a continent to another. Because of this, filling in the gaps associated with unconformities is often impossible. What is required, other than the absolute timescale afforded by radioisotopic dating, which has been available only in the last half-century, is some sort of indexing system for rock layers. Such an indexing system exists, in the form of fossils.

Defined broadly, a fossil is evidence in Earth's rocks of an organism (plant or animal) that once was alive. Fossil remains fall broadly into one of six categories:

1. actual remains, including mummified bodies, sharks' teeth, etc.
2. petrifications, in which the original organic matter has been completely or partially replaced by mineral matter such as calcite, quartz, chert, pyrite, or others
3. molds or casts of the inside or outside of bodies
4. prints of leaves or of soft-bodied creatures such as jellyfish, and including detectable amounts of unidentified remnant organic matter
5. fossilized excrement
6. tracks, trails, and burrows of animals.

Examples of fossils are shown in Figure 8.5. The processes that result in fossilization are varied, depending on the type of fossil being considered. For tracks or imprints, the sedimentary template must not be modified greatly as the loose grains of sand, mud, or pebbles slowly cement together under modest compression to form a sedimentary rock. Metamorphism, except for the lowest grade variety, will distort such prints out of existence.

Petrification of biological structures is the sort of fossil with which most people are familiar, and those of large animals make up the most spectacular museum displays. The formation of such fossils depends on the chemical similarity between the elements carbon, the primary element in the dead organism after desiccation, and silicon, a predominant element in both continental and oceanic rocks. Examine the periodic table in Figure 2.6, and note that carbon and silicon are in the same column of the table, occupying adjacent rows. As discussed in Chapter 2, the significance is that carbon and silicon require the same number of electrons to complete or close a valence "shell"; the particular ways in which these elements combine with others therefore are similar. Because carbon and silicon occupy a central column of the table, many possible bonding combinations are possible; this complexity leads to some differences in their chemical behavior.

A dead organism resting in sediments, and eventually covered by them, exists in a universe that is predominantly rock. Hence there is far more silicon than carbon in the environment of the remains. In any physical system, atoms and molecules are continually exchanging places with other compatible atoms and molecules. Usually (considering now only atoms), atoms will

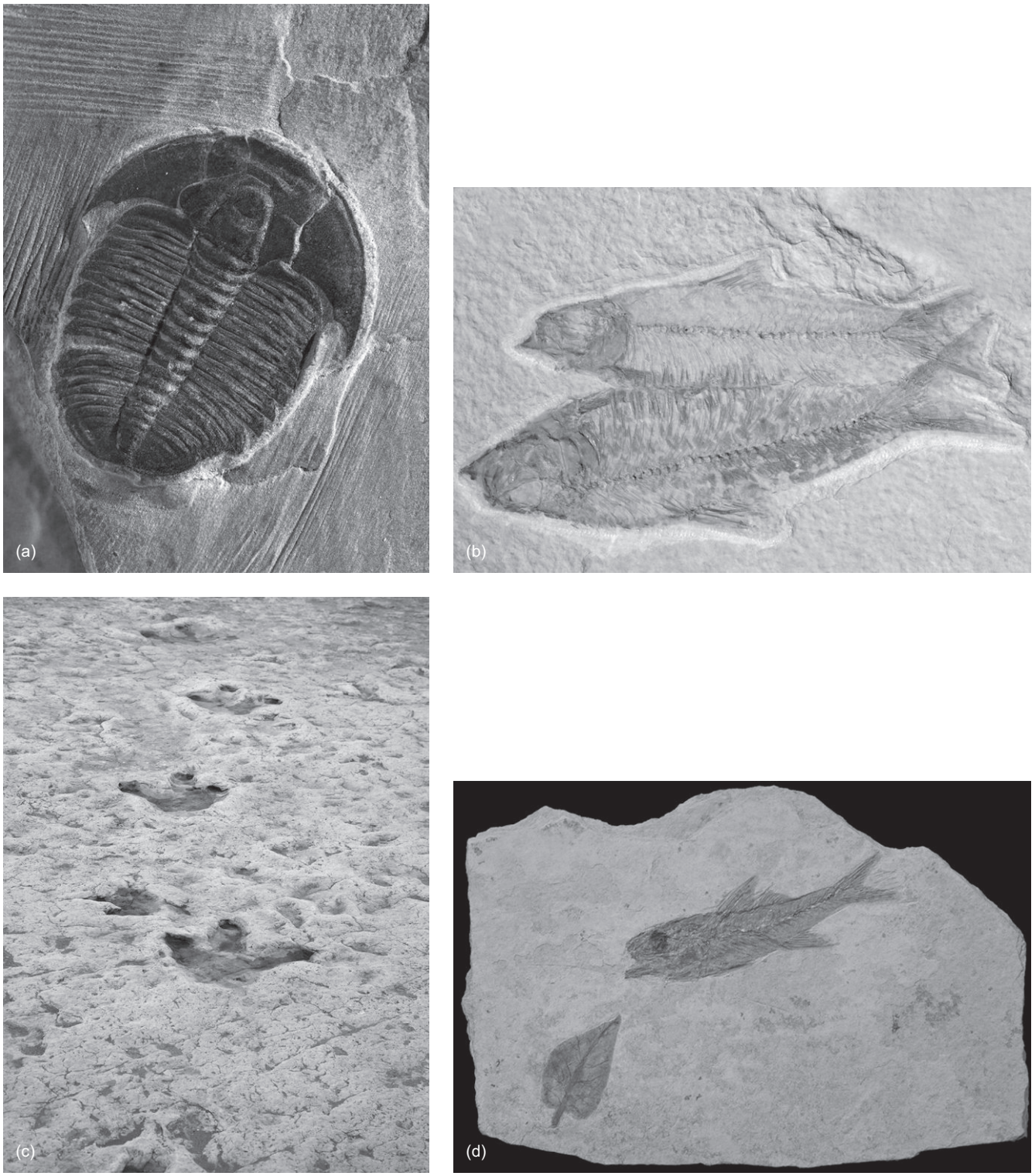


Figure 8.5 Examples of fossils: (a) *Elrathii kingi* trilobite from middle Cambrian period, © Russell Shively; (b) *Knightsia* species fish from Eocene epoch, © Jubal Harshaw; (c) dinosaur tracks from the Jurassic period, near Tuba City, Arizona, © Mark Higgins; (d) fish with leaf from Cambrian period, © Marcel Clemens. All images from Shutterstock.com.

exchange with other atoms of the same element. However, if a chemically similar element is present in far greater numbers, it may substitute for the atoms of the original element. This is the case with organic remains in sediments: gradually, the carbon is replaced by silicon from the surrounding material. If the sediments are undisturbed, so that the location of the substituted carbon is preserved, the silicon atoms reproduce the structure of the organism. The outlines of the fossil are discernible because its texture and overall chemical composition are distinct from that of the rocky grains around it, even though both contain silicon.

The process of natural fossilization is a chancy and rare event. The vast majority of individual organisms that lived on this planet show no evidence of their existence, their atoms being consumed by other organisms or otherwise dispersed into various natural chemical systems and not replaced point by point with silicon. However, those creatures that have been fossilized, from the microscopic to the gigantic, provide a record of what life existed at the time a particular sedimentary layer formed. Because life has changed over the history of Earth – new forms replacing older forms and whole ecosystems changing in a clearly recognizable way – fossils provide an index of the particular time when sediments were laid down. This timescale is still relative because no absolute measure of dates is available, but it can be correlated from one place to another on Earth. Unconformities can be identified when sedimentary layers with a particular fossil sequence are present in a geologic column in one part of the world but not in another.

Index fossils identify sedimentary rock formations as having occurred at a certain time in the history of Earth. Because fossils are not found in igneous rocks, and only rarely in metamorphic rocks, such rocks must be fitted into the time sequence according to their relationship to sedimentary layers that they may overlie, underlie, or intrude.

8.6 Radioisotopic dating of Earth rocks

The fitting of igneous rocks into the sequence defined by indexed sedimentary layers provides a mechanism for absolute dating of such layers, by radioisotopic dating of the igneous rock. The radioactive clock in an igneous rock begins ticking when the rock solidifies from a melt. Before that, the atoms in the liquid magma are so mobile that daughter elements of a radioactive decay will migrate away from their site of formation, making dating impossible. Thus the age of a terrestrial rock is the time since it last solidified. Metamorphic rocks may be dated *radiometrically*, that is, by radioisotopes, but the age derived is ambiguous. It may reflect the time of metamorphism, if it were of the high-grade variety, but more often the atoms have only partially rearranged themselves, leading to an apparent age that is meaningless. Sedimentary rocks themselves cannot be dated radiometrically; the age thus derived reflects when a particular grain was originally solidified in an igneous rock, or may be uninterpretable if the grain came from a metamorphic assemblage. The time at which the grains were eroded from the rock outcrop and transported by water is not reflected at all in the isotopic age.

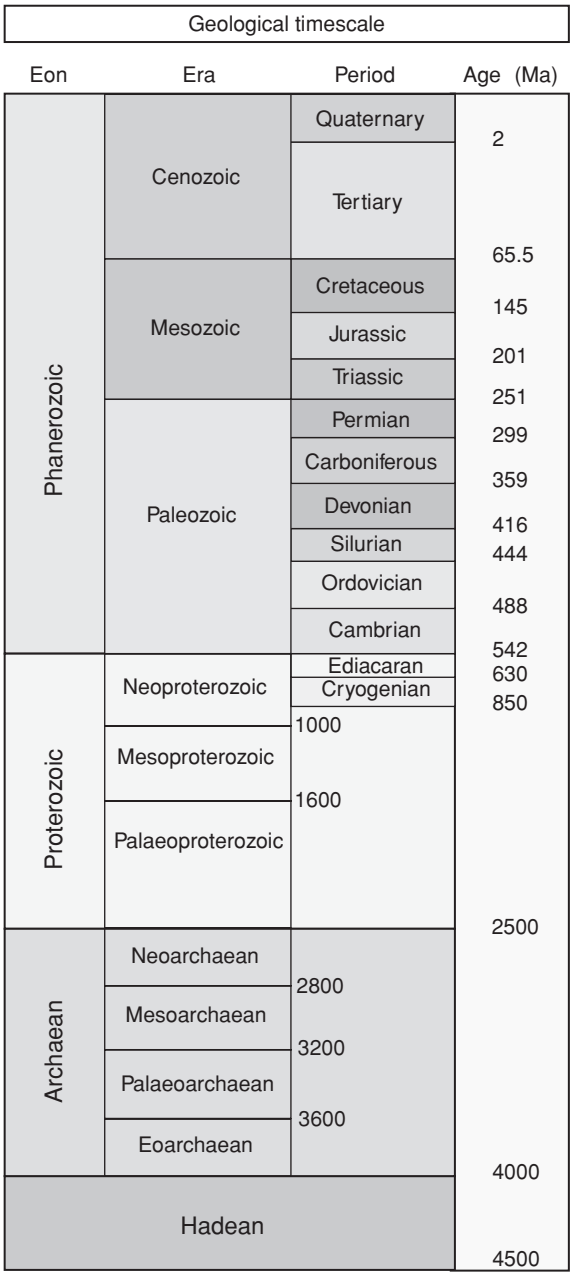


Figure 8.6 The geologic timescale, divided into eons, eras, and periods, with absolute times in millions of years on the right. Figure from Johnson and Harley (2012), by permission of Cambridge University Press.

8.7 Geologic timescale

The relative timetable assembled from sedimentary sequences was divided by geologists according to major changes in the types of rocks present, and the appearance or disappearance of groups of fossils from layer to layer. Some boundaries in the timetable correspond to the apparently sudden extinction of a significant fraction of Earth's species in existence at the time. We discuss in Chapter 18 very recent insights into the origins of such extinctions.

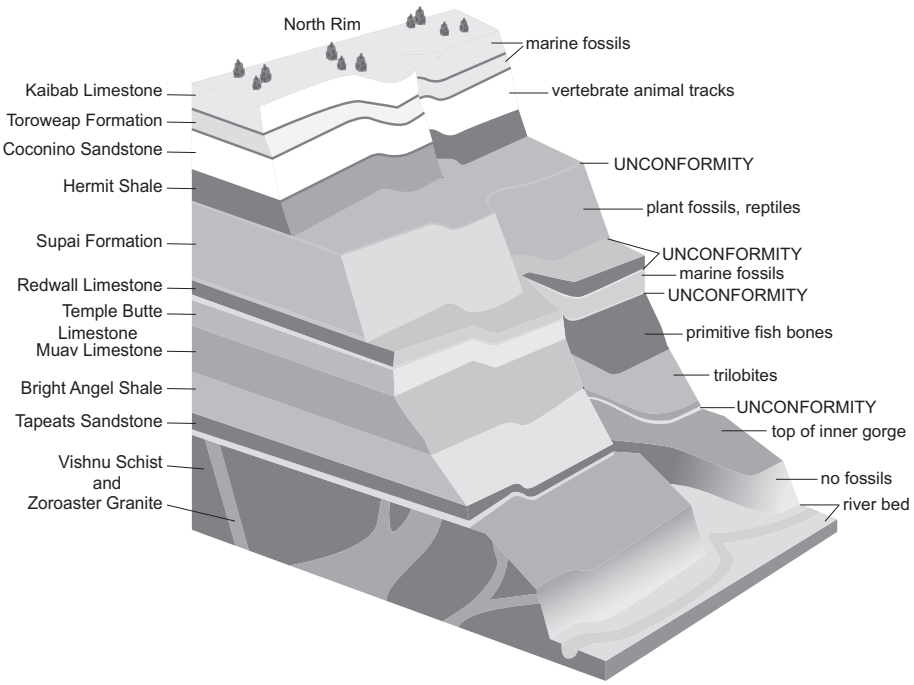


Figure 8.7 Sketch of the sedimentary layers exposed to view along the South Rim of the Grand Canyon of the Colorado River in Arizona. Unconformities and some key fossil types are indicated.

There is little regularity to the nomenclature associated with the geologic timetable (Figure 8.6). Layers often are named after particular regions in which the first, or most famous, of a particular sequence of rocks is found. For example, the Jurassic period is named after an exposure of limestone rocks in the Jura mountains of Europe. A hierarchy of divisions is employed in naming rock layers: *eon, era, period, epoch*, and *age*, from largest to smallest divisions. Eons are divided according to major transitions in the rock record. The transition between Priscoan and Archean at 4.0 billion years marks the appearance of the oldest rocks on Earth; Archean to Proterozoic at 2.5 billion years was set originally at the age of the oldest discernible fossil *stromatolites*, remains of certain types of bacterial colonies, earlier fossil evidence of life now has been found. The Proterozoic to Phanerozoic transition at 570 million years (Ma) roughly marks a rapid diversification and complexification in the types of lifeforms on Earth.

Eras are finer divisions more closely associated with the appearance of certain fossil groups, and came into use before the division of time into eons. The Paleozoic is the first era of the Phanerozoic eon. The Paleozoic–Mesozoic boundary is set at the beginning of the dominance of dinosaurs in the fossil record; the Mesozoic to Cenozoic similarly marks the sudden disappearance of dinosaurs and proliferation of mammals. Periods are even smaller time divisions, and date back to attempts in the nineteenth century to divide the history of Earth, like a play, into three acts: Primary, Secondary, and Tertiary, with only the last (youngest) name surviving in current usage. Finally, epochs and ages represent small groups of, and individual, stratigraphic units for which variations in nomenclature from continent to continent are allowed.

8.8 A grand sequence

The classical example of a sedimentary sequence is that of the *Grand Canyon of the Colorado*, which cuts through the high plateau country of northern Arizona. From the South Rim to the Colorado River is a vertical drop of about 1.6 km (1 mile), and more from the North Rim. The canyon was formed as the Colorado River cut through the landscape, which was uplifted into a plateau sometime over the past few tens of millions of years. However, the sequence of rocks through which the river cuts covers a time period from 230 million to 1.7 billion years ago. The sequence consists primarily of sedimentary layers, with metamorphic rocks and then igneous rocks toward the bottom. The sequence reflects not just deposition of material, but periods of erosion as well; unconformities are present, the largest of which represents a missing 600 million years (or 13%) of Earth history. That time is not lost: elsewhere in the region, including parts of the North Rim, and farther afield various portions of the missing sequence are present.

Figure 8.7 sketches the Grand Canyon sequence. Most of the various rock layers are named after aspects of the history of the Grand Canyon and the region. A given rock layer will correspond in time to many others around the world; those other layers will differ depending on the particular environment present in each locale at the time. A beach environment will look different in the rock record than a lake bottom, ocean bottom, or mountain region. Names of rock formations identify the location and hence the particular kind of layer present; in each locale, these can be tied to a particular time in the geologic timescale by relative or absolute dating techniques.

8.9 The geologic timescale as a map

The geologic timescale of Figure 8.6 is a kind of road map into the past, based on hundreds of years of work by geologists in the field mapping out layers of rock, determining the rock types and kinds of fossils if sedimentary, correlating with radiogenic dates on associated igneous rocks, and tying the sequences together worldwide. Our planet is not the only one with a geologic history – all of the planets and their moons have one. Characteristic of Earth, however, is the predominance of sedimentary processes based on transport of material by water. Only Mars is likely to have a similar (but much less extensive) form of depositional history – one can see evidence for layering in the sides of the major Martian canyon, Valles Marineris, and elsewhere. Whether fossils are present on Mars, though, is an open issue that we address in Chapter 15. And we cannot rule out sedimentary processes on other worlds where the liquid medium is not water: deposits of solid organics laid down by running liquid methane are an exotica that might exist on Saturn's largest moon, Titan.

Summary

The Earth's landforms record the geologic forces that have been at work upon them, but interpretation is difficult because of the variety of processes and the incomplete nature of the preservation of the record. In the last few decades it has become clear that many changes are relatively sudden on geologic timescales, even catastrophic; an example is hypervelocity impact of asteroid fragments. Other processes, such as erosion of mountain chains, may occur gradually over long timescales. Traditional estimates of the age of the Earth based on the accumulation of sediments eroded from mountains fall short by a large factor because they cannot account for large numbers of missing sedimentary layers in one particular region – a so-called "unconformity". Rocks can be considered to begin their "lives" in the form of lava or magma, molten material extruded from or injected into the Earth's crust, which then solidifies. Some igneous rocks are exposed when the entire region into which they were intruded is uplifted, and then erosion by wind and water removes the weaker rock around it. Eroded material may remain in lowlands as sand dunes or "alluvium", or be transported down to lakes and the ocean by streams and rivers. Sedimentary material will cement together to form rocks such as sandstones; the increased overburden will subject older layers to progressively higher temperatures and pressures, chemically transforming it into so-called metamorphic rock. This may be uplifted and exposed at the surface by erosion, sometimes forming part of a melange of material with igneous and sedimentary rocks as well. Other metamorphic and sedimentary rocks may be so deeply buried that they are fully remelted and reappear as igneous rocks. Fossils, the traces of once-living organisms, provide an index by which relative dating of different sedimentary layers can be obtained. Fossils include actual organic remains, lithified forms in which silicon has replaced the original carbon, tracks and trails of animals, and molds or casts of bodies. Radioisotopic dating of igneous intrusions into sedimentary rocks has allowed absolute dates to be assigned to the relative sequence of geologic time indexed by fossils: the so-called geologic timescale. Among the most extensive sequence of sedimentary layers, providing a glimpse into more than a billion years of history, can be seen in the Grand Canyon of Arizona in the United States.

Questions

1. Speculate on the political and cultural reasons why the enormous age of Earth was not recognized until the eighteenth century.
2. Consult a geologic map for the area in which you live to infer some of its geologic history. What was it like 100 million years ago where you live? 10 million years ago?
3. While a sedimentary or metamorphic rock that has been remelted in the crust to make an igneous melt loses its original physical properties, does it lose all trace of its prior chemical composition? Justify why a chemical fingerprint might remain, and what the implications are for repeated cycling of rocks from igneous to sedimentary and metamorphic, and back to igneous.
4. In the chapter the formation of metamorphic rocks was described in terms of altered sedimentary rocks. Can igneous rocks be metamorphosed?

General reading

Cloud, P. 1988. *Oasis in Space: Earth History from the Beginning*. W. W. Norton, New York.

References

Breed, W. J., Stefanic, V., and Billingsly, G. H. 1986. *Geologic Guide to the Bright Angel Trail: Grand Canyon, Arizona.* American Association of Petroleum Geologists, Tulsa, OK.

Johnson, M. R. W. and Harley, S. L. 2012. *Orogenesis: The Making of Mountains*. Cambridge: Cambridge University Press.

Mason, S. F. 1991. *Chemical Evolution*. Clarendon Press, Oxford.

Murray, O. 1986. Greek historians. In *The Oxford History of the Classical World* (J. Boardman, J. Griffin, and O. Murray, eds). Oxford University Press, Oxford, pp. 186–203.

Press, F. and Siever, R. 1978. *Earth 2/E*. W. H. Freeman and Company, San Francisco.

Stevenson, D. J. Nature 2009.

Wyllie, P. 1971. *The Dynamic Earth*. John Wiley and Sons, Inc., New York.

9

Plate tectonics: an introduction to the process

Introduction

We close the part of the book on techniques for discerning Earth's history with a conceptual tool. The concept of *plate tectonics*, whereby the outer layer of Earth is divided into a small number of distinct segments called *plates*, which move relative to each other, represents a breakthrough in explaining a diverse range of geologic phenomena across our planet. Although the basic ideas are now 30 years old or more, this picture or concept of how Earth's geology works, in a unified way, continues to provide fresh insights into evolution of Earth, the stability of the gross climate of our planet, and the distinctions between Earth and the other planets. Because of its importance, we introduce the concept early to allow the reader to gain an understanding of the basic ideas. We come back to plate tectonics again and again as a fundamental process on Earth driving climate change, erosional processes, atmospheric chemistry, and even the nature of life.

9.1 Early evidence for and historical development of plate tectonics

Revolutions in scientific thinking often take place when increasing numbers of observations challenge existing theories, which in many cases have become dogmatic over time in the face of conflicting data. Particularly satisfying is the synthesis of widely diverse data into a single framework that explains well all of the data.

As early as Sir Francis Bacon over 350 years ago, but mostly since the early nineteenth century when maps of the world became good enough to reveal the true shapes of the continents, the significance of the curious matching of the edges of distant continents has been pondered. Africa and South America seem to fit fairly well into each other, and those people inclined toward jigsaw puzzles found they could put North American, Europe, Africa, and Asia into a plausible single landmass (Figure 9.1). What this implied for the origin of today's continents, that somehow they broke apart from one or several bigger landmasses, seemed ludicrous to most geologists used to working on a rather solid Earth.

Other lines of evidence caused trouble for static models of Earth's continents. In the nineteenth century, it was noticed that fossil organisms in a given region, particularly on southern continents, reflected ecosystems with widely varying climates, ranging from cold desiccated regimes to lush tropical jungles. Some scientists argued that Earth has tumbled through its history, shifting the positions of the high latitudes (i.e., poles) through the various continents over time. However, also noticed was that, prior to the Cretaceous period of Earth history, similar or identical fossil plants and animals were found on continents separated today by deep oceans. During and after the Cretaceous, these biological linkages largely disappeared. The idea of temporary land bridges was advanced to allow nonmarine organisms to transfer between continents. However, it was evident to some that allowing continents to split apart and drift away from some common original supercontinent could explain both types of observations. Further support came from the observation that North and South American mountain ranges that terminate at the Atlantic Ocean line up very well with ranges in Europe and Africa, as if they had once been continuous belts within the inlands of a single larger continent.

In the late nineteenth century, catastrophic models were proposed to provide mechanisms for breaking up and moving continents. The idea was advanced that Earth's Moon was originally material blown out of Earth by asteroidal collision from what is now the Pacific Ocean, and the resulting stresses on a single continent on the other side of the world caused fracturing that separated the continents and opened the Atlantic Ocean. (As we see in Chapter 11, current models of lunar origin are not too dissimilar from this basic concept, though very different in the details.) The most famous proponent of continental drift, German meteorologist Alfred Wegener, suggested in 1915 that,

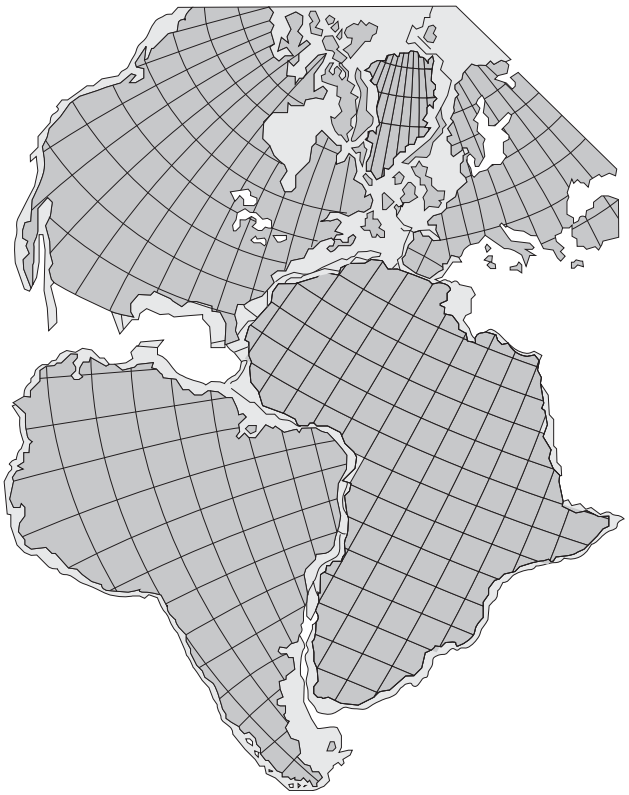


Figure 9.1 Modern version of a jigsaw puzzle worked since the 1600s: the fitting together of the continents around the Atlantic Ocean. The solution shown was done by computer to yield the least amount of overlap. The gridded portion of continents shows land currently exposed; the outlined but ungridded area is *continental shelf*, which is the part of the continental landmass that is currently underwater. (For this fitting exercise, only that part of the shelf extending down about a kilometer below sea level was taken.) The fit is not perfect, but many of the misfits arise from portions of continental material that turn out to be younger, that is, added later than the time when the continents are thought to have split apart. Modified from Wyllie (1971) by permission of John Wiley and Sons, Inc.

because Earth is slightly oblate, with a larger radius at the equator, continents might pull apart from a single continent originally centered at the poles. He also argued that a westward drift might be caused by the gravitational pull of the Sun and the Moon. He saw the continents, made mostly of granite, as being able to plow through the basaltic crust of the oceans.

Wegener's enthusiasm for continental drift did not play well in conservative scientific circles, and in fact, the gravitational tidal forces he invoked to induce drift are much too weak to do so. As a result, the concept gained the flavor of a crackpot model in mainstream scientific circles, and helped delay the realization and acceptance of plate tectonics by four decades.

9.2 Genesis of plate tectonics after World War II

9.2.1 Seafloor topography

The decades after World War II saw a burst of activity in many sciences, including geophysics. Techniques were developed – in large part from military *sonar* technology, which uses echoes of sound waves to locate underwater structures – to map the ocean floor's topography, that is, the distribution of elevation caused by undersea mountains and trenches, among other features. The results were striking, as shown in Figure 9.2. The ocean floor is subdivided by long ridges of mountains, stretching over thousands of kilometers. These ridges are cut transversely by a series of fractures, or *faults*. Near the edges of some continents or major island chains, long trenches are present, the deepest of which is the Marianas trench extending down 11 km from the surface of the ocean, some 7 km below the average seafloor depth.

Even more striking is the distribution of height over the surface of Earth, shown in Figure 9.3. Essentially, Earth has two different kinds of surfaces, characterized by their depth: continents and oceans. The range of depths within each kind of surface is generally less than the difference in depth between them, leading to a *bimodal distribution*. As we see in Chapter 15, this is dramatically different from the situation on Mars and Venus, suggesting that Earth's geology has been shaped by a global set of processes unique to our planet. Whether plate tectonics is in fact unique to Earth is explored critically in Chapter 16.

The mid-ocean ridges and trenches are distinguished not merely by their topography, but also by the fact that earthquakes and volcanoes are concentrated along their lengths (Figure 9.4). Earthquakes also are concentrated along certain fault systems, such as the famous San Andreas fault of California. Because earthquakes are the result of stresses built up by movements in the outer layers of Earth, the presence of narrow belts of activity suggest that the crust is characterized by organized motions rather than by random distortions or deformations.

9.2.2 Magnetic imprints on rocks

Further evidence for moving continents came from an entirely different field of study, called *paleomagnetism*. The magnetic force is one expression of the general electromagnetic force discussed in Chapter 3, which is manifested by charged particles that are in motion. Certain materials, such as the mineral *magnetite*, exhibit the ability to retain a permanent magnetization associated with the alignment of the spins of the electrons in the atoms of which they are comprised. Such a magnetic force will cause tiny scraps of iron (iron filings) to line up in a particular manner, which defines the direction of the *magnetic field* of the grains in the mineral (Figure 9.5).

Where does the magnetic field of magnetic or, more precisely, magnetizable, minerals come from? Anyone who has played with a compass is familiar with Earth's own magnetic field, which, as with a bar-shaped magnet, has a definite directionality to it. A magnetized iron needle suspended so as to rotate freely (i.e., a compass) will point in the general north–south direction – not quite due north–south, because Earth's magnetic field is not precisely aligned with the geographic poles. The origin of Earth's magnetic field remains an outstanding puzzle in planetary sciences and is discussed further in Chapter 11.

Important to us here is that, as minerals crystallize from molten rock, they do not immediately become capable of holding a magnetic field. This is because at high temperatures the electrons are too mobile to acquire a permanent fixed direction to their spins. Instead, the mineral grains must pass below a

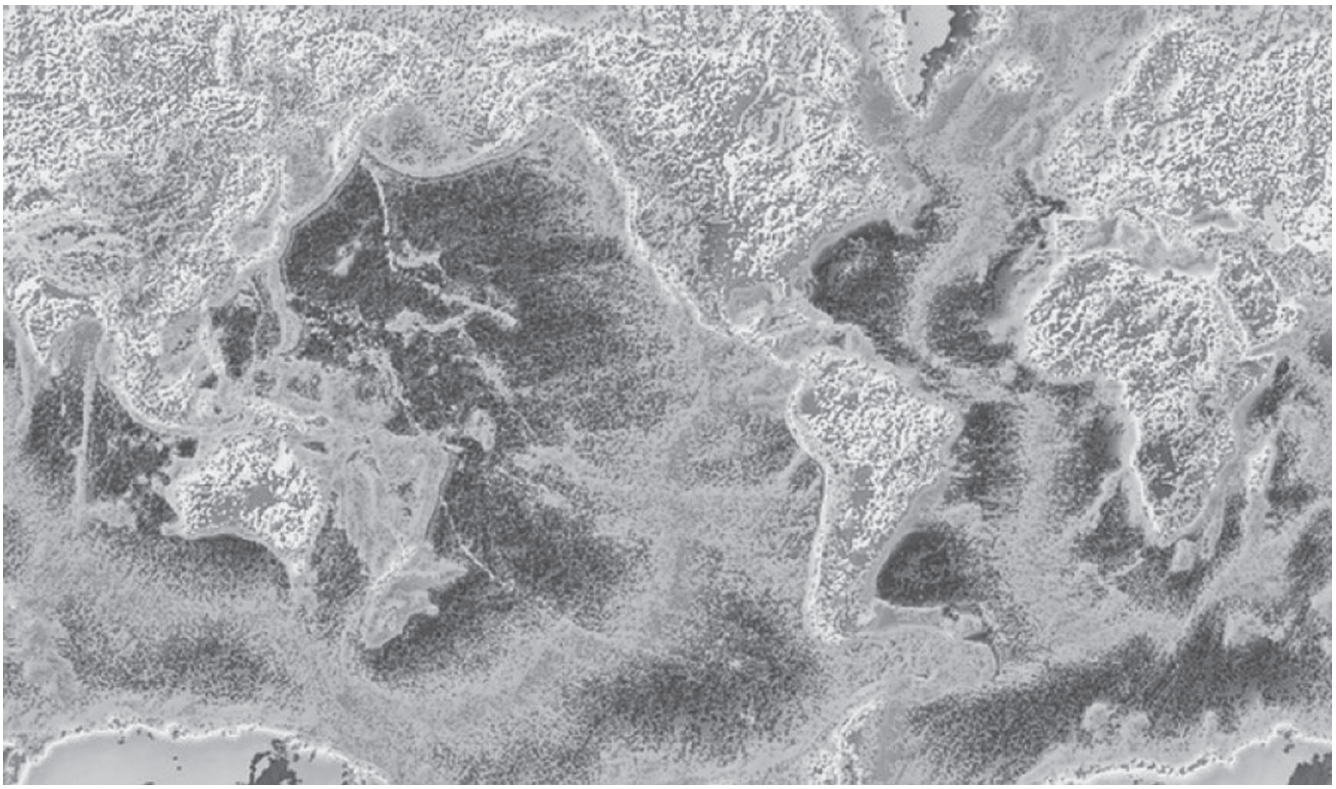


Figure 9.2 Map of the Earth's topography. Beneath the oceans, the seafloor topography is dominated by vast mid-ocean ridges, transform faults, and subduction zones, as described in the text. Red is highest, and blue is lowest, elevation. See color version in plates section.

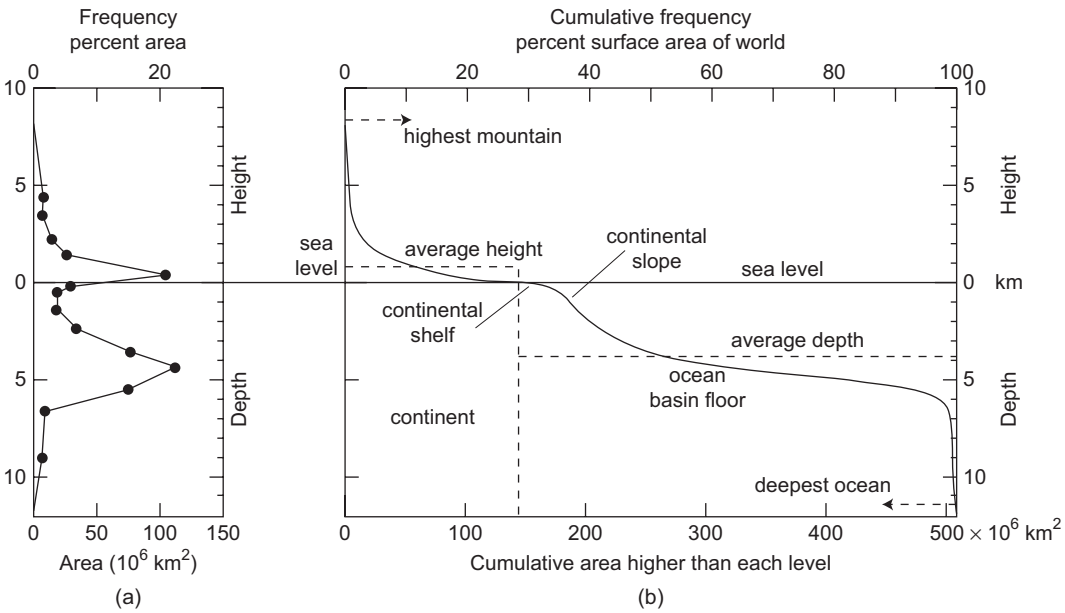


Figure 9.3 Amount of area at various heights and depths on Earth, relative to sea level. (a) Amount is expressed in terms of total area occupied by surfaces with a given height (or depth) in kilometers. The bimodal nature of Earth's topography is evident. (b) The same data are expressed in terms of cumulative area higher than a given altitude; this gives a sense of the gross profile of continents and ocean floors. From Wyllie (1971) by permission of John Wiley and Sons, Inc.

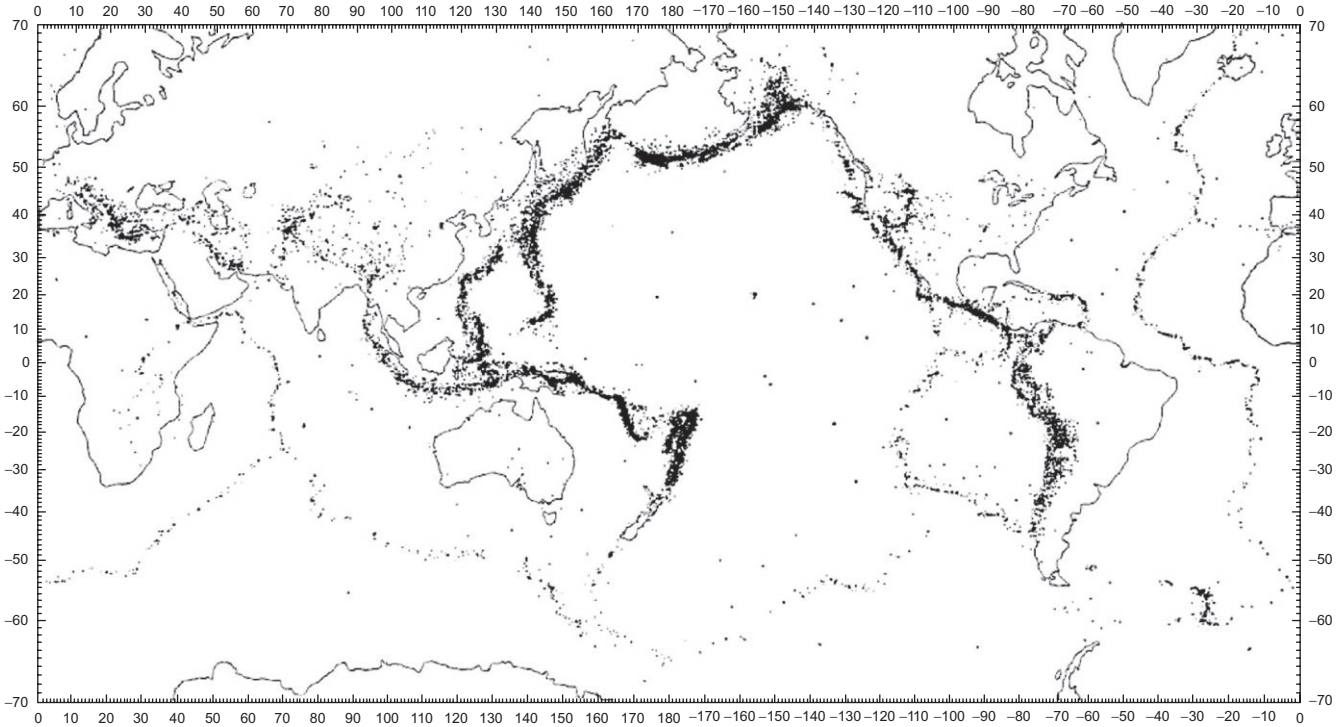


Figure 9.4 Map of earthquake activity over the surface of Earth. Each dot represents a single epicenter. The geographic pattern of volcanic eruptions is very similar. From Isacks *et al.* (1968).

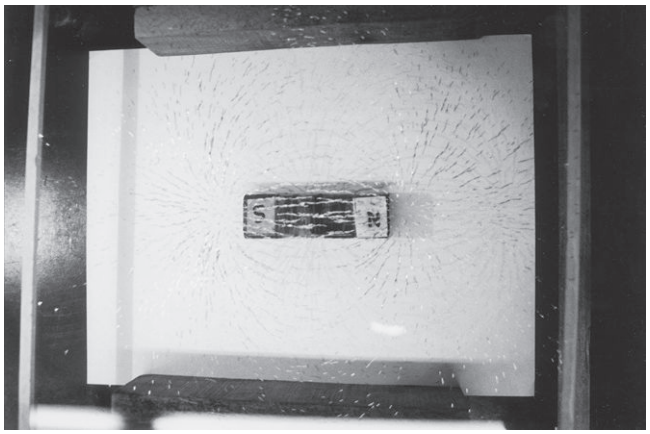


Figure 9.5 Appearance of iron filings when a permanent magnet, in this case bar shaped, is placed underneath the glass upon which they rest. Lab set up courtesy of Larry Hoffman, University of Arizona Physics Department.

temperature known as the *Curie point* to retain a permanent magnetization. This temperature is roughly 800 K and is well below the crystallization point at which the minerals solidify from the melt. Because of this, the direction of the magnetic field inside a permanently magnetic mineral depends on the orientation of the mineral, relative to Earth's magnetic field, at the time it cools through its Curie point.

Hence, if a mineral cools and magnetizes, it becomes a kind of compass. If the rock formation in which the mineral is embedded is rotated by 90 degrees, the "north pole" of the magnetic field in the mineral will point roughly east or west, not north. This would seem to provide an excellent means for determining the movement of continents relative to Earth's magnetic field, provided that one can date the rocks by radioisotopic techniques. However, interpretation of this information up through the 1950s often focused on the idea that the rotational poles of Earth had shifted with time – either that Earth slowly tumbles through space, or that the entire outer layer of Earth slips over the interior in one piece. In this interpretation, there is no relative drift of the continents but only synchronous changes in latitude around the world.

Studies of continental rocks indicated that rotations in the apparent field directions could not be accounted for solely, or even primarily, by a coherent shifting of all of the continents. Instead, continental drift had to be invoked to reconcile the directions of remnant magnetization in rocks of various ages and locations. But how were the continents moving relative to each other? The key observation came from the direction of magnetization of rocks on the seafloor.

Study of continental rocks revealed that the Earth's magnetic field has reversed direction many times in the past – a compass held fixed during a reversal would swing from south to north, or vice versa. These reversals could not be accounted for by a 180-degree rotation of the rock itself, because rocks of the same age from a variety of locations and orientations show the same reversal. The origin of the reversals lies in the way Earth's magnetic field is generated, deep in its interior, but is as yet very poorly understood. Regardless of their origins, magnetic reversals provide yet another way of delineating the progression of geologic time in the rock record, provided the ages of the

reversals are determined by independent dating of rocks, for example, by radioisotopes.

In the 1950s, technologies developed to detect submarines by their magnetic signatures began to be employed to map the magnetic orientations of seafloor rocks from surface ships. Very quickly, a remarkable regularity emerged at the mid-ocean ridges, illustrated in Figure 9.6. Stripes of differing magnetic field intensity are laid out parallel to the ridge itself. The variations in intensity are most straightforwardly interpreted to be caused by the direction of Earth's magnetic field at the time the rock cooled from a magma. The pattern mimics that of the field reversal history recorded in rocks on the continents, indicating that the youngest rocks are closest to the mid-ocean ridge, increasing in age farther out. The simplest and most straightforward interpretation of the pattern was advanced in the 1960s: the ridges are sites where new ocean crust is being created, moving like a conveyor belt away from the ridges. As the new crust cools after extrusion, the field direction is recorded in magnetic minerals as they cool below their Curie point. This interpretation quickly led to another question: if new seafloor is being created at ridges, where is it being destroyed?

9.2.3 Geologic record on land

By the 1960s it was becoming increasingly difficult to refute the notion that continents moved about Earth, and that seafloor spreading was somehow involved. Geologic patterns on continents were now being re-evaluated in light of the apparent mobility of Earth's surface. Radioisotopic dating of similar types of igneous rocks on the eastern end of South America and western end of Africa showed a remarkable correspondence – a well-delineated boundary separating rock of 2-billion-year and 600-million-year ages in western Africa was present in eastern South America as well, and in just the right place for a good jigsaw puzzle fit.

Rocks in mountain ranges in northern California, on the west side of the very active San Andreas fault, matched up well in type and age with rocks a couple of hundred kilometers to the south, in southern California, on the *east* side of the fault line. It was more than tempting to simply slide the east and west portions of California, along the fault, so that these rock types matched. Measurements spanning many decades of ground slippage resulting from earthquakes showed a northward movement on the western side of the fault of roughly a centimeter per year. Using that figure, the now-separated rock formations would have been together some tens of millions of years ago.

Drift timescales of centimeters per year seemed to fit in other parts of the world. The magnetic anomaly pattern on the mid-Atlantic ridge displayed ocean floor ages consistent with spreading of a few centimeters per year. Studies of the Hawaiian island chain showed that the age of the islands increases to the northwest, with active volcanism confined to the Big Island (Hawaii) on the southeast end and a still-growing submerged island (*seamount*) just to the southeast of it. If interpreted in terms of the ocean crust drifting northwestward over a source of volcanism, drift rates of some centimeters per year again are obtained.

With mounting evidence in the 1960s for continental drifting and seafloor spreading, the question of how the growing crust was accommodated was answered through the study of earthquakes.

9.2.4 Earthquakes and subduction

Networks of *seismometers*, which measure the local shaking of the ground due to near and distant earthquakes, are capable of inferring both the geographic location (epicenter) and the approximate depth at which the sudden shifting of rock occurred which caused the quake. (More on this is given in Chapter 11.) Most earthquakes occur at shallow depths beneath Earth's surface. Careful comparison of Figures 9.3 and 9.4 shows that quakes are common at mid-ocean ridges, trenches, and at sites of lateral movement such as the San Andreas fault in California. However, earthquakes that originate at depths greater than 100 km are confined to the trenches, where some quakes originate as deep as 600 km below Earth's surface (Figure 9.7).

Earthquakes generally occur where stresses build up in rocks and that stress is relieved suddenly by a failure or fracturing of the rock. Stress build up is generally the result of some movement that forces rocks up against each other; in zones such as the San Andreas, such motion is a sliding one in which rocks lock up against each other and then break free. In the trenches, a map of earthquake locations showed an interesting pattern: the deeper quakes were actually displaced, usually toward the continental side of the trenches (Figure 9.7). The most natural interpretation is that the trenches are regions where ocean floor is sinking underneath continents; as the movement occurs, lock up of rocks causes stress build up and eventual release through fracturing. Thus the trenches are sites at which ocean crust is destroyed: the other end of the conveyor belt that begins at the mid-ocean ridges.

With understanding of the nature of the great seafloor trenches, the basic picture of floating continents was completed in the 1960s and became the theory of *plate tectonics*. Tectonics refers to the study of movement and deformation of the outermost layer of Earth, its crust; plate refers to the emerging concept that Earth's crust was divided into discrete plates. Seafloor is created at plate boundaries called mid-ocean ridges and destroyed at *subduction zones*, which are the ocean trenches. Buoyant continental crust rides passively on the plates, buckling but (for the most part) not subducting when continents on two converging plates collide. This relatively simple model tied together a wealth of geologic data accumulated over the first half of the twentieth century and has shaped our thinking about the geology of Earth and other planets since then.

9.3 The basic model of plate tectonics

The Earth's crust is broken up into a small number of relatively rigid plates that move slowly across the surface in response to forces generated beneath the crust, in the *mantle*. Strictly speaking, the terms crust and mantle refer to chemical differences between the layers. It is more accurate to refer to the plates as comprising the Earth's *lithosphere* (rock-sphere), a rigid outer shell that rides on a hotter and plastic (not molten) layer called the *asthenosphere* (weak-sphere). The crust–mantle boundary

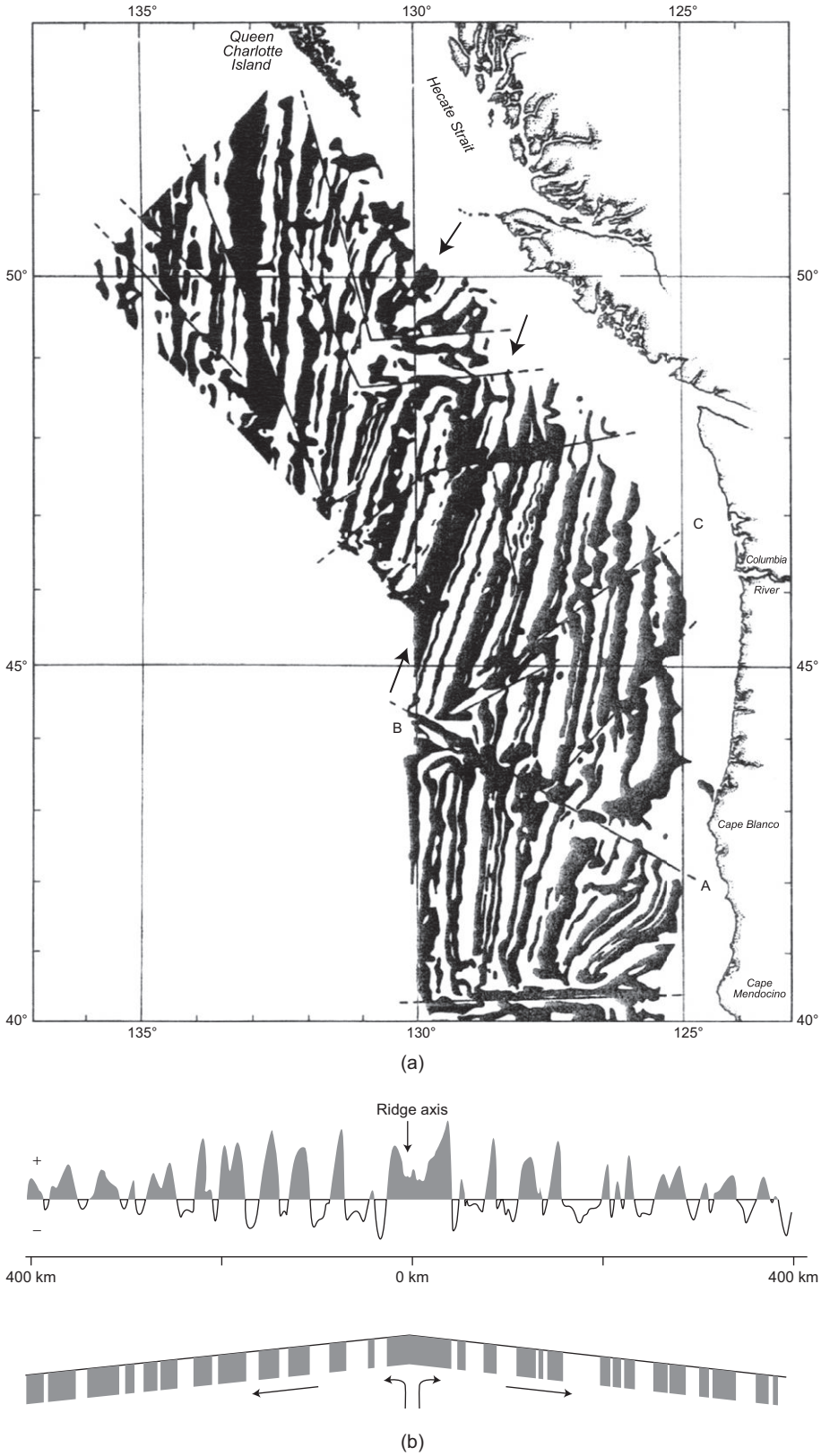


Figure 9.6 Magnetic record of seafloor spreading: (a) Magnetic anomaly pattern on the axis of the Juan de Fuca ridge, near Vancouver island. Black indicates current magnetic field direction, white the reversed field. From Morgan (1968). (b) Interpretation of the cause of magnetic reversals, shown as a cross-section through the top of the Earth's oceanic crust. Arrows indicate the sense of spreading. On the upper chart, the axis of the mid-ocean ridge is marked, and the relative strength of the magnetic field as a function of distance from the axis is given. Positive (gray) is the present direction of the field. From van Andel (1992).

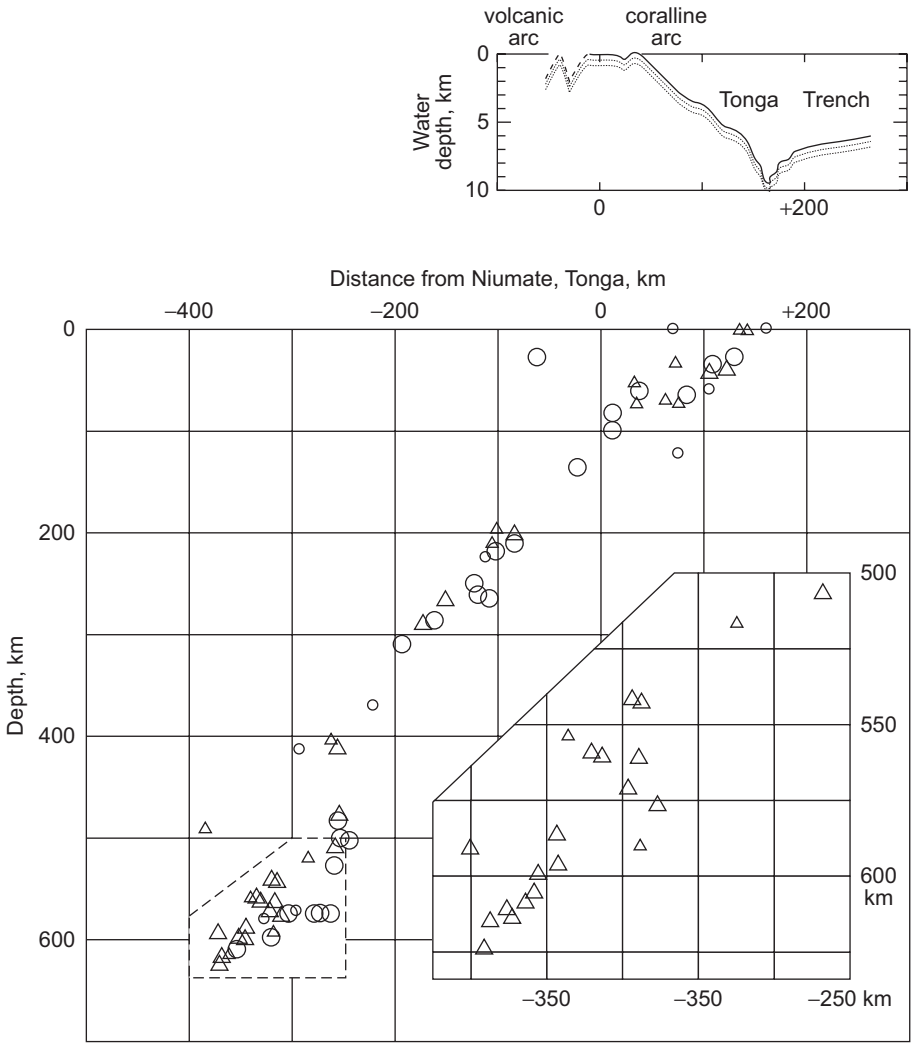


Figure 9.7 Earthquake location and depth near the Tonga subduction region in the Pacific. A profile of the topography (exaggerated 13 times) is shown in the upper small panel, positioned so that the zero point is aligned with that on the big figure. Note that the position of the quakes with depth seems to outline quite nicely a subducting slab of crust moving diagonally under Earth's surface. The symbols distinguish between earthquakes that occurred north of the station at Niumate, Tonga, (circles) and those occurring south of Niumate (triangles). The inset shows the deepest quake locations in more detail. From Isacks *et al*. (1968).

is not the same as the lithosphere–asthenosphere boundary, but there is significant overlap between the mantle and the asthenosphere. As discussed in Chapter 11, knowledge of the presence of a plastic layer beneath a rigid outer shell comes from observing the slowing of earthquake waves, which must pass through this layer; the chemical differences are inferred from material erupted to the surface in certain volcanoes.

At the mid-ocean ridges, hot buoyant magma rises to the surface, cools, and freezes, forming new seafloor. As this seafloor moves laterally away from the ridges, it cools and eventually becomes dense enough to sink, forming a subduction zone. The sinking slab differs from the mid-ocean ridge material not only because of its temperature and hence density; it contains rock that has reacted chemically with seawater, as well as sediments delivered by rivers from the continents as well as the remains of countless shell-forming organisms. Some of the sediments are scraped off in the shallow part of the subduction zone, but some survive to deeper levels. The sinking slab is subjected to increasingly higher temperatures and pressures as it descends, to the point where at least some of it is assimilated into the surrounding mantle material. How much of the slab is assimilated remains an active topic of debate: some computer models of subduction indicate that the majority of slab material stays cool enough to sink to the base of the mantle, forming a "graveyard" of sunken slabs. Such debris might eventually absorb enough heat from the decay of its own radioactive elements, as well as from the underlying core, to rise again as the "mantle plumes" that are responsible for the Hawaiian Islands and other island chains within plates. Other models, with different assumptions about the details of the chemistry and heat transfer, produce nearly complete assimilation of the slabs within the mantle. The release of water from the heated slab triggers melting of overlying mantle material, leading to volcanism as described in Chapter 16.

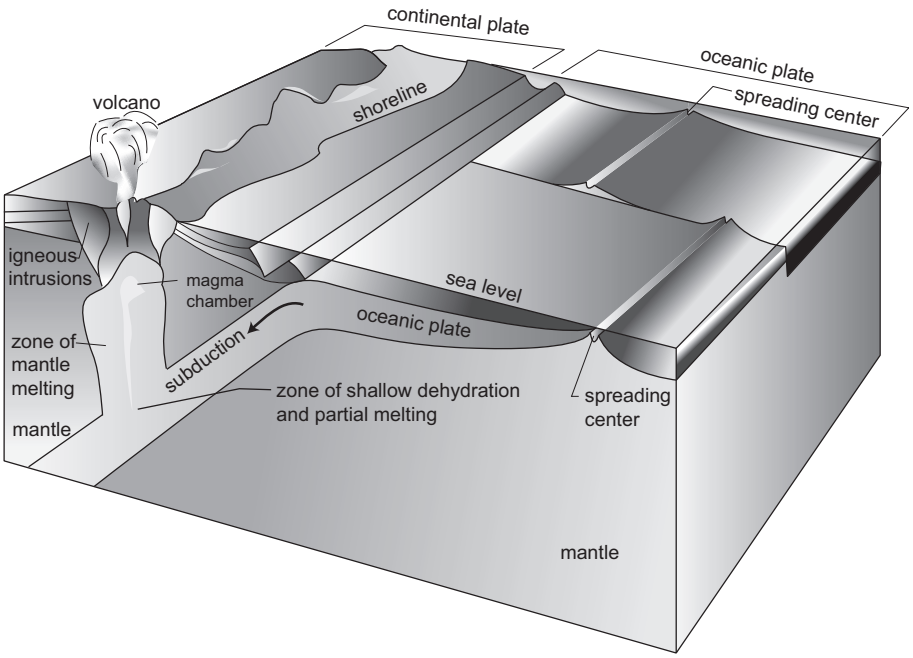


Figure 9.8 Oblique view of Earth's lithosphere illustrating the motion at mid-ocean ridges (spreading centers) and subduction zones. The spreading center is shown split by a transform fault.

Although subduction can occur at the boundary between two ocean plates, most subduction zones are on the margins of continents, because continental crust is buoyant compared to ocean crust. The density difference is chemical in nature, ocean crust being basalt, which is a relatively dense rock compared to continental granites. The origin of this difference, one of the key enablers of plate tectonics and a distinguishing feature of Earth compared with the other planets, is considered in Chapter 16. Thus, ocean crust rides somewhat like rafts on a "sea" of oceanic crust, overriding the oceanic crust at subduction zones. Volcanism in the interior of a continent may then occur as heated slab material releases water to the upper mantle, melting it (Chapter 16). Other types of interactions, such as continent-continent collisions, or "obduction" in which a small piece of continental crust breaks off and is shoved into the subduction zone, with ocean crust riding over it and welding onto the edge of the continental boundary.

Figure 9.8 illustrates the nature of the basic plate boundaries. Ridges and trenches where crust is brought up and slides back down, respectively, are not sufficient to accommodate plate motion on a spherical Earth with irregular continents. Instead, the ridges are sliced through by *transform* faults, where lateral motion occurs. The San Andreas fault is part of a transform system, which connects the eastward-moving subduction of the Pacific plate on the west coast of South America with the northerly and northwesterly subduction occurring along western Canada across to Asia. A simplified map of ridges, trenches, and transform faults around the world is given in Figure 9.9.

The speeds at which plates move range from 1 to 10 centimeters per year, corresponding to a thousand kilometers in 10 million to 100 million years. Aside from the geologic evidence that tells us, among other things, how long ago certain well-separated rock formations were together, and magnetic reversal stripes on the seafloor that give us the rate from the calibrated history of reversals, the space age provided direct measurement capability. Astronauts placed reflectors on the surface of the Moon, which geologically appears to lack plates and to be a rigid surface. Bouncing laser beams off the Moon from various continents on Earth allows the relative movements of the stations to be determined directly. More recently, Earth-orbiting satellites that calibrate their position by making accurate observations of deep-sky objects have been able to make similar direct determinations of relative movements of ground stations. Finally, large radio telescopes on Earth that also determine their position by staring at the sky, and are linked together by computers, can determine continental drift as well. The measured speeds are similar to inferred speeds at which the very plastic rock in the asthenosphere of Earth slowly turns over, removing heat to the surface.

Current plate motions and associated geology are complex (Figure 9.9). The concentration of earthquake and volcanic activity along the Pacific Rim is the result of the northwestward movement of the Pacific plate, where it is taken up by subduction along the Aleutian islands and the east coast of Asia. Southward from Indonesia, the Pacific plate adjoins the Australian plate, which generally is moving northward. Although the plates are approximately rigid bodies, collision between continents on separate plates leads to compression and uplift of mountains or plateaus. The collision of India on the Australian plate with Asia is the most dramatic example of this, pushing the Himalayan plateau up to the greatest altitudes above sea level anywhere on the planet.

On the east side of the Pacific, the plate's northward motion is expressed by transverse sliding of the plate along the North American plate. The San Andreas fault system is where this

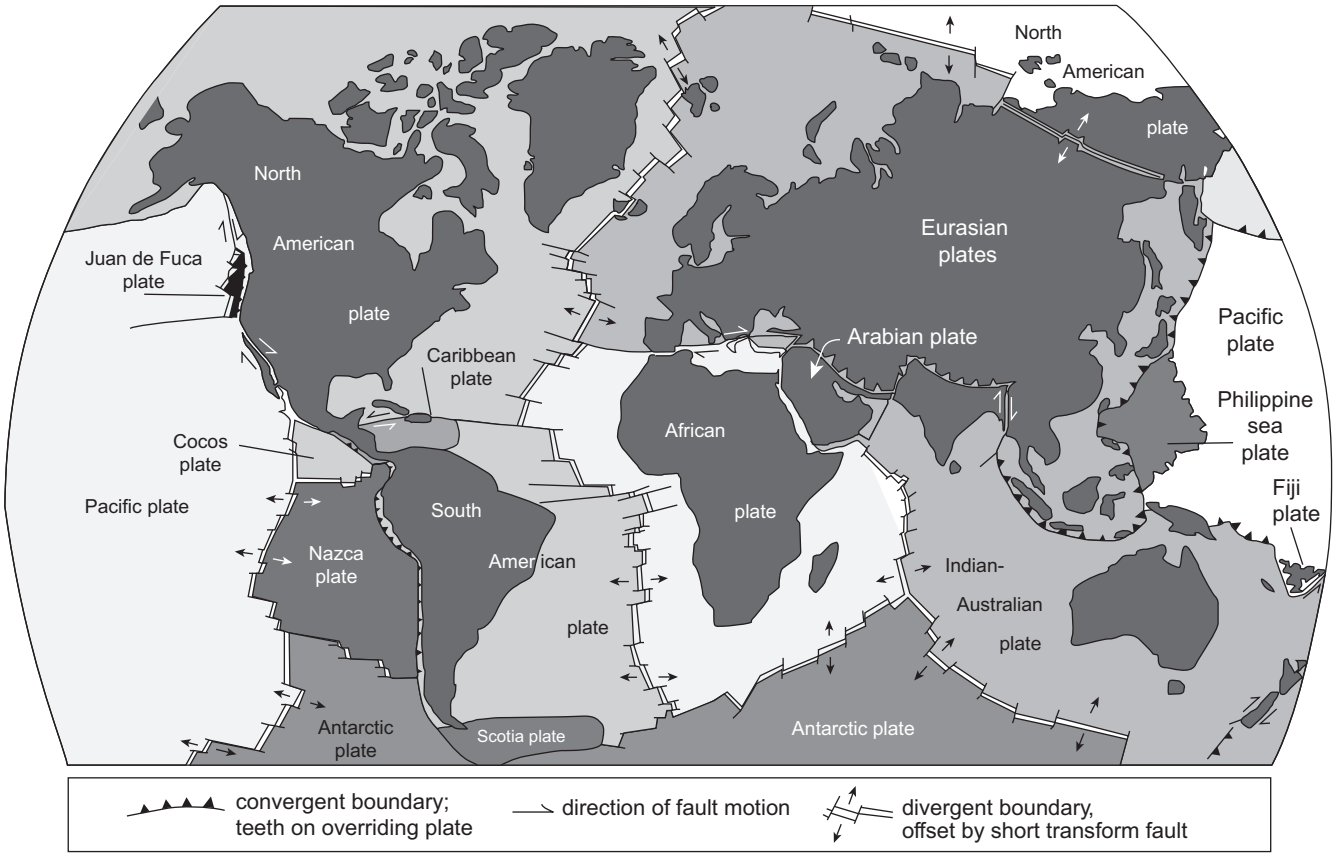


Figure 9.9 Schematic map of Earth's system of ridges, trenches, and transform faults. Plates and the general directions of their motions are labeled. Adapted from Cloud (1988) by permission of W. W. Norton Company.

motion is accommodated on the continent, leading to the prodigious earthquake activity in California. Farther north, the Pacific plate undergoes subduction, the expression of which includes volcanoes such as Mount St. Helens. The material brought upward from the heating of subducting plates includes basalts, rhyolites (a volcanic form of granitic rocks), and a kind of hybrid between basaltic and granitic compositions, called *andesite*. These volcanic products reflect a complex sequence of mixing of oceanic crust with scraped-off detritus of the underlying portions of the adjacent continent, followed by melting and eventual ascent to the surface.

Farther south, subduction of the Cocos and Nazca plates under South America is expressed in extensive mountain building and volcanic activity along the Andes. Spreading between the Cocos and Nazca plates, and between the Pacific plate and both of these smaller plates, creates a complex *triple junction*, the motion of which is accommodated by the subduction and transverse motions along the western end of North America.

The situation in the Atlantic is simpler and quieter. Spreading of that ocean along the mid-Atlantic ridge, dividing the African and Eurasian plates to the east and the South American and North American plates to the west, continues. The African plate is also rotating in such a way that the African continent is moving northward, pushing the floor of the Mediterranean Sea into southern Europe and building the Alps. The Arabian landmass is moving into Asia, and a zone of spreading between Africa and Asia includes the Arabian Sea and very substantial rift valleys on the east coast of Africa itself.

9.4 Past motions of the plates and supercontinents

Reconstructing the positions of the continents back in the past is straightforward for recent times, because one simply has to take today's plate motions and run them in reverse. Eventually, however, one reaches a point at which the current continents have amalgamated into two continents, and then even earlier, into one supercontinent. At that point, roughly 200 million years ago, plate motions cease, and the earlier history of plate tectonics must be deduced from similarities between ancient rock formations on different continents, as well as the magnetic orientations of rocks.

Figure 9.10 shows the history of spreading from 200 million years ago (Jurassic times) to the present. *Pangaea*, a name derived from the Greek pan (all) and gaia (earth), was a single Jurassic supercontinent comprising all current landmasses. It was surrounded by a global ocean called Panthalassa "all sea". It began to break up with North America and Eurasia forming *Laurasia*, from the Greek laura, meaning passage or channel, and the remaining continents comprising *Gondwana*, named by

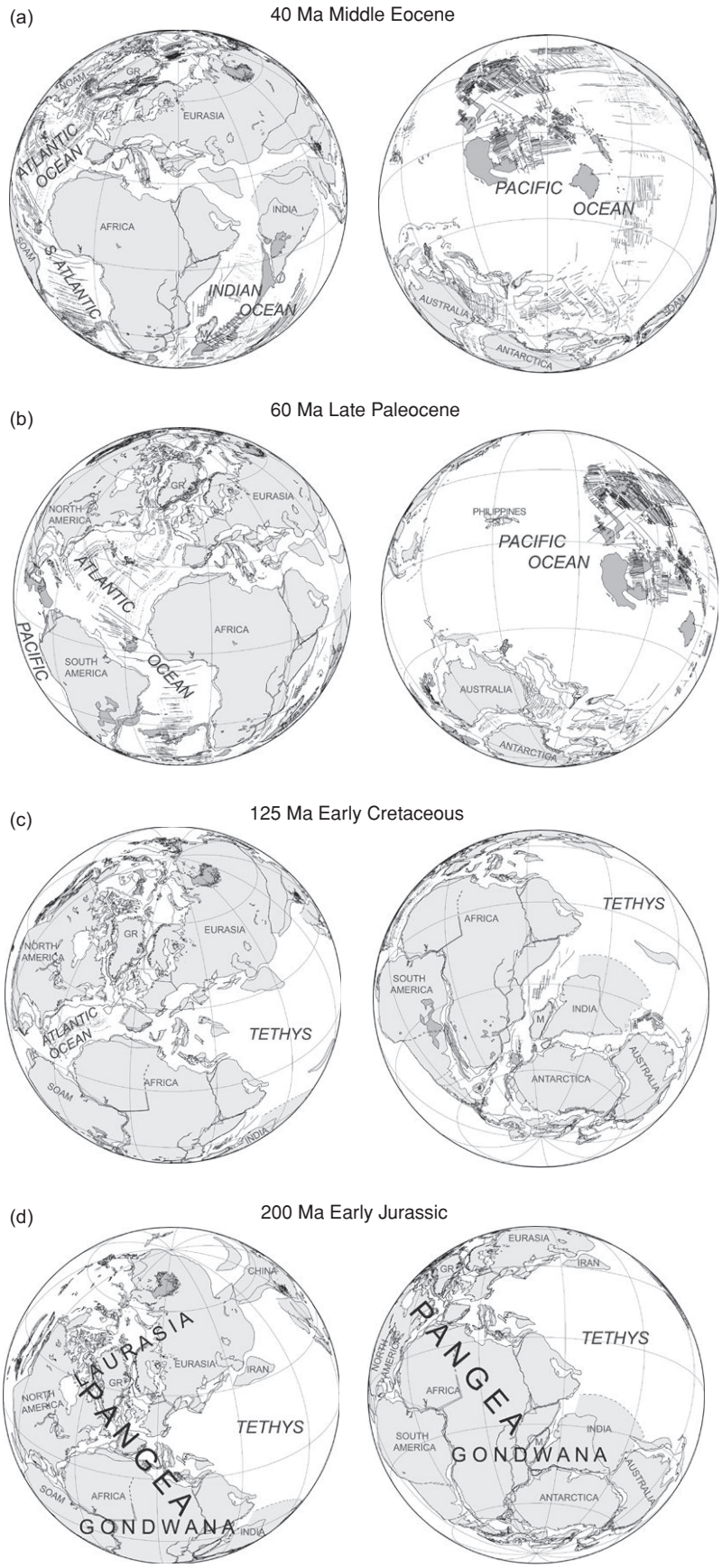


Figure 9.10 A depiction of the movement of the continents back to (a) 40 million years ago; (b) 60 million years ago; (c) 125 million years ago; and (d) 200 million years ago.

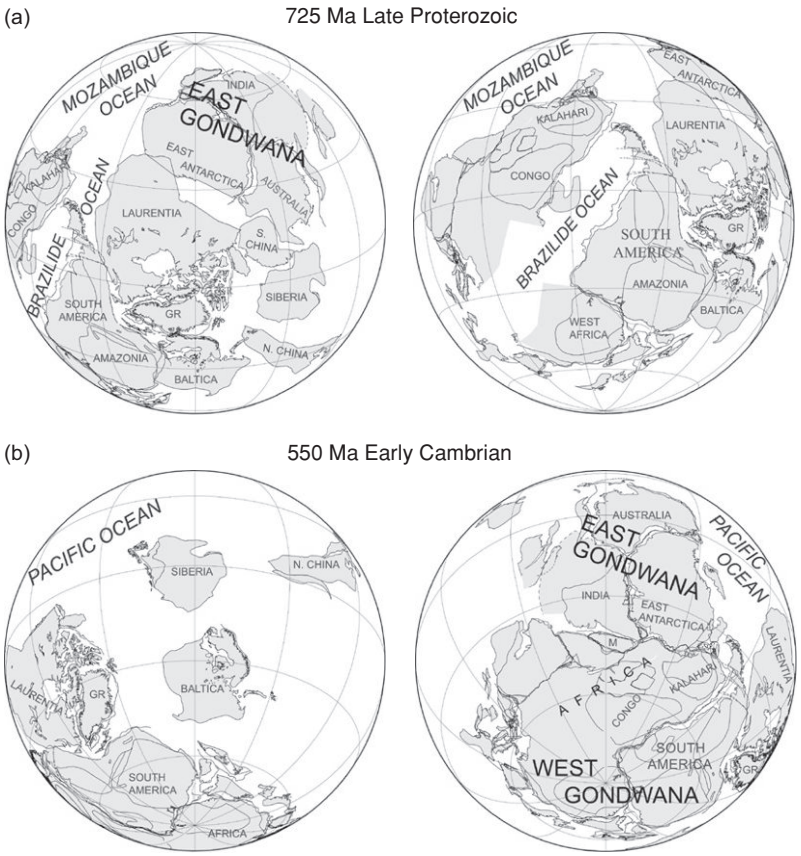


Figure 9.11 Continental drift from (a) 725 million to (b) 550 million years ago, showing the supercontinent previous to Pangea shown in Figure 9.10.

a contemporary of Wegener, for unknown reasons, for an ancient tribe in India. The *Tethys* seaway between the two is the ancestral Mediterranean, meaning mid-land, Sea. By 135 million years ago, Gondwana had begun to break up into Africa/South America, Antarctica/Australia, and India. At the close of the Cretaceous, some 65 million years ago, Africa had reconnected with Eurasia, but the major new event was the opening of the Atlantic Ocean, the growth of which continues at present. Extrapolating 50 million years into the future, the Atlantic will widen, a new sea will open in East Africa, Australia will cross the equator, the present southern California coast will shift north of San Francisco and the Gulf of California will merge into the Pacific.

While this scenario generally has been considered sound for about three decades, less certain is the history of the continents prior to 200 million years ago. Detailed detective work on rocks and their magnetic orientations provide some clues. Rocks from prior to the Mesozoic era in North America and Africa have different magnetic directions, such that they could not be pieced back together into a single Pangaea prior to about 260 million years ago. Evidence showing that ancient ocean basins actually closed to form Pangaea at this time comes from volcanic rocks of ocean crust composition appearing in mountain ranges of Argentina, Africa, and parts of the Appalachians, these rocks being forced up onto the continents by the closure.

Reconstructing the continental configurations prior to Pangaea raises an important point: because of the continuous process of subduction of ocean crust, there are few seafloor rocks older than roughly 250 million years before present. On the continents, which do not subduct, rocks dating back billions of years are not uncommon. Here and there on the continents, very ancient oceanic crust can be found, where it was pushed upward and sutured onto continents during collisions. These preserved remains of ocean floor, or *ophiolite suites*, stand in stark contrast to the vast bulk of the oceanic crust that is part of a youthful and dynamic conveyor belt of material moving from ridge to trench in a matter of some hundred million years or so.

Expeditions to Antarctica have found sandstones in the mountains of that continent from about 700 million years ago that contain fossil remains of worms identical to those in rocks of similar age in Wisconsin and other locations around the world. This and other resemblances in widely separated rocks of the era push one toward the existence of another supercontinent, named Rodinia, in existence from perhaps as long as 2.2 billion years ago. The break up of that continent about 700 to 800 million years ago and eventual reassembly into Pangaea some half-billion years later illustrates the cyclic nature of the formation and break up of supercontinents. (Most reconstructions imply an intermediate supercontinent, Pannotia, came together very briefly about 500 million years ago.) Figure 9.11 illustrates

the possible sequence of the break up of Rodinia and formation of Pangaea, based on computer simulations that account for the rock relationships seen on the continents of today.

Prior to Rodinia the record of continental motions is extremely limited. The fossil record is very poor before 750 million years ago, in part because the variety and abundance of preservable organisms was impoverished prior to that time. The explosion of new and complex biological types is a remarkable watershed about 600 million years before present that we consider in Chapter 18. The rock record itself is spottier, both fossils and magnetic signatures being more poorly preserved because of subsequent modification of the rocks. Furthermore, the decreasing abundance of rocks at earlier and earlier ages is not simply a matter of preservation: the continents themselves have grown over time, and prior to several billion years ago, much of the landmass we see today did not exist (Chapter 16). Thus the history of plate tectonics prior to a billion years before present, including assembly and disruption of supercontinents, is not well understood.

What is now understood from the efforts to track the motions of continents over hundreds of millions of years is that creation and destruction of supercontinents must have had important effects on the climate and biota of Earth. This is not simply due to the changing latitudes of individual localities but more profoundly to the global effects on ocean currents and landmass ice sheets of having continents and oceans in different configurations. We discuss these possible effects in Chapter 19.

9.5 Driving forces of plate motions

What drives the motion of the plates? To understand fully the origin of plate tectonics requires thinking about how Earth gets rid of its heat, a subject covered more fully in Chapter 11. However, some important aspects of the forces behind plate motion can be considered here.

The mechanics of plate motion suggest two mechanisms for moving plates. At mid-ocean ridges, hot material is welling up from the asthenosphere/mantle to form new crust. This hot material must be less dense than its surroundings in the mantle, in order for it to rise. As it reaches the surface it must move laterally away from the ridge, creating a pushing force as more material rises to take its place. This *ridge push* has been invoked by some as a means to drive spreading motions. At subduction zones, or trenches, oceanic crust is forced under continents back into the mantle. For this process to take place, oceanic crust must be denser than continental, and indeed, the oceanic basalts are significantly denser than the continental granites and other rocks. However, for subduction of the slab to continue downward some 600 km or more (based on earthquake data), the slab also must be denser than the surrounding mantle. How can ocean crust be buoyant relative to mantle at ridges while not being so at the trenches? Cooling of the ocean crust leads to a sufficient increase in density of the rock that it sinks into the mantle. This process is aided by the thinness of the ocean crust, which allows it to be easily bent and redirected into the mantle at the trenches.

A very crude analogue of this process is seen at volcanic lava lakes. Hot material rides to the surface, cools, solidifies, and gets dense enough to sink. In the case of ocean crust and mantle, the basalt within a ridge is molten, but it solidifies quickly as it rises out of the ridge. The density increase that allows sinking occurs in the solid basalt as it moves away from the ridge and cools further.

The effect of slabs of ocean crust sinking is to produce a pulling motion on the plates. This force also has been invoked as that which sets the plates in motion. In a sense, one can think of the downward sinking slab as a cantilevered beam, a beam supported on one end only. But in this case the beam is not being held fixed by a rigidly anchored support; instead, it is dragging the support – the entire oceanic plate – along with it. Hence any point on the plate is moving toward the the subduction zone, and the slab is merely that part of the plate that has cooled sufficiently to become denser than the underlying asthenosphere and sink.

Detailed modeling of both slab-pull and ridge-push suggests that the pulling action is most important. However, this cannot be the case when the continents are all combined into a single supercontinent, which has happened at least twice in Earth's history. The initiation of continental break up must be caused by a phenomenon other than slab-pull. What has been suggested is that the supercontinent, having stalled the motions of the plates, tends to prevent heat from escaping from Earth's interior. Hot spots in the mantle therefore form, which move buoyant material rapidly upward to thin the base of the supercontinent, causing rifting and eventual break up. An alternative is that the hot spots do not originate because of the presence of the supercontinent, but arise very deep in Earth for unknown reasons, and those appearing under supercontinents are the ones that happen to cause their break up.

What is the origin of the chemical differences between mantle, oceanic crust, and continental crust? Armed with an understanding of the basic plate tectonic mechanisms, we discuss this issue in Chapter 16. It is a critical one, because the subtle density differences between these different types of rock enable plate tectonics to occur, and these differences in part depend on the presence of liquid water. The continued stability of liquid water on Earth over time is dependent on a climate regime that may have been moderated by plate tectonics, leading to a tightly coupled and complex relationship between water and plate motions.

It is valid and worthwhile to ask whether other planets and moons in the solar system possess plate tectonics. The answer seems to be no for the other terrestrial planets. Mars has a thick crust and cool interior, which long ago stalled plate tectonics. Venus, so similar in size to Earth, has a surface characterized by massive volcanic flows and little or no evidence for plate tectonics. It seems the lack of water and very high surface temperatures have played a role in preventing Earth-type tectonics from starting there, both because water lubricates the plates and changes the mechanical behavior of the rock itself. We delve deeper into the reasons for Earth being unique in regard to plate tectonics in Chapter 16.

For those who were educated in geology prior to the 1950s, the processes shaping Earth seemed to be vertical ones: mountains get built upward, and subsidence creates ocean basins. Plate tectonics turned geologic thinking sideways: the primary motions are horizontal, but in the compression and break up of plates lie the origin of most mountain ranges and new seas. Throughout

the book we see profound effects of cycling crust over time, as important atmospheric gases get captured in ocean sediment and cycled beneath continental crust, only to be released again. The interplay between plate tectonics and Earth's atmosphere and biosphere is one of the surprises of late twentieth century geology and planetary sciences, one with a sobering impact on the possibilities for stable climates equable for life elsewhere in the cosmos.

9.6 An end to techniques and the start of history

In this part of the book we have considered tools and concepts essential for our examination of the history of Earth and other planets. We turn next to that history, starting at the beginning of our solar system and its central star, the Sun.

Summary

Plate tectonics is a description of the way in which the Earth's crust moves in response to the heating of the crust from below. The development of plate tectonics is an excellent example of how diverse kinds of evidence can challenge existing theories and force their replacement by new, initially controversial ones. Although the shapes of the continents and the correspondence of fossil species long suggested the idea that they might have once been joined together, only after the topography of the seafloor was revealed, along with the presence of stripes of alternating magnetic polarity revealing the presence of seafloor spreading, was the idea taken seriously. It is along the mid-oceanic ridges that new seafloor is produced from the rising and solidification of magma, leading to the spreading of the seafloor. As the seafloor moves away from the spreading centers, it cools and becomes denser. Eventually the cooling seafloor becomes denser than the underlying mantle on which the crust rides, and it founders and sinks in the form of subducting slabs. The zones of subduction corresponding to deep-sea trenches, usually (but not always) found along the margins of continents. The continents represent a different type of crust, less dense and hence more buoyant than oceanic crust, which with only minor exceptions is never subducted into the mantle. The overall movement of the plates on a spherical Earth also requires places where plates are moving laterally relative to each other. The patterns of earthquakes and volcanism are strongly correlated with the edges of plates, and different types of volcanism occurring at spreading centers, mid-ocean ridges, and in the interiors of plates reveal the various processes by which melting occurs associated with plate tectonics. Among Earth, Venus, and Mars, only the Earth seems to have plate tectonics today, although Venus might have experienced it in the past.

Questions

1. Why do you suppose geologists of the early twentieth century were so reluctant to consider continents moving across the globe, in view of the fact that they accepted as plausible large *vertical* movements?
2. The technologies available to geologists after World War II provide an excellent example of how military technology can create scientific revolutions. What other areas of science or medicine were revolutionized as a result of military developments in World War II?
3. How is the geology of the area where you live related to the global pattern of plate tectonics? Has the geologic nature of your area been determined mostly by subduction, by spreading, by transform motion, or by hot spot volcanism (for example)?
4. In what ways does the lateral motion of plates depend on the Earth being a sphere; that is, how would plate tectonics be different were the Earth a flat sheet in which the plates were small compared to the size of the sheet?

General reading

Cloud, P. 1988. *Oasis in Space: Earth History from the Beginning*. W. W. Norton, New York.

National Research Council. 2008. *Origin and Evolution of the Earth*. The National Academies Press, Washington DC.

References

Browne, M. W. 1995. Experts ponder causes of breakup of ancient supercontinent. *New York Times*, Oct. 3, p. B5.

Cloud, P. 1988. *Oasis in Space: Earth History from the Beginning*. W. W. Norton, New York.

Dalziel, I. W. D. 1997. Neoproterozoic–Paleozoic geography and tectonics: review, hypothesis, environmental speculation. *Geological Society of America Bulletin* **109**, 16–42.

Gutierrez-Alonso, G., Fernández-Suárez, J., Weil, A. B. *et al*. 2010. Self-subduction of the Pangaean global plate. *Nature Geoscience* **1**, 549–53.

Isacks, B., Oliver, J., and Sykes, L. R. 1968. Seismology and the new global tectonics. *Journal of Geophysical Research* **73**, 5,855–99.

Morgan, W. J. 1968. Rises, trenches, great faults and crustal blocks. *Journal of Geophysical Research* **73**, 1,959–82.

Smith, W. H. F. and Sandwell, D. T. 1997. Global sea floor topography from satellite altimetry and ship depth soundings. *Science* **277**, 1956–62.

van Andel, T. H. 1992. Seafloor spreading and plate tectonics. In *Understanding the Earth: A New Synthesis* (G. C. Brown, C. J. Hawkesworth, and R. C. L. Wilson, eds). Cambridge University Press, Cambridge, UK, pp. 167–86.

Wyllie, P. 1971. *The Dynamic Earth*. John Wiley and Sons, Inc., New York.

PART III

The historical planet: Earth and solar system through time

10

Formation of the solar system

Introduction

Having dealt with some of the tools and key concepts to which we will return as we develop the history of Earth and the other planets, we are ready now to consider that history. Five centuries after the beginning of the European Renaissance, humanity's explorations of Earth and the cosmos have exposed an intriguing, perhaps profound, paradox. Earth and the other planets of the solar system seem to be explainable as manifestations of common physical processes that have operated over very small and very large scales, to produce a range of cosmic phenomena. In this sense we are neither special nor particularly important in the grand scheme of things.

On the other hand, in our own solar system, we now understand Earth as the one planet with a uniquely stable climate on its surface, equable for liquid water over the billions of years required to bring forth intelligent life. Although Mars may have come close to this state at one time, the surface appears lifeless today. Europa may have a habitable oceanic environment beneath its icy crust. An intriguing possibility is that Saturn's moon Titan may have had a stable "hydrosphere" over its history, but one in which methane substitutes for water: whether such an environment could be habitable for a very exotic form of life is not known (Chapter 12). Other solar systems may be common and life may flourish elsewhere, but it is also possible, with what we know today, that we are a rare or even unique speck in the cosmos. We will know more over the decades to come, but for now we seek to understand how this planet came to be, and how physical processes have operated to make it habitable for billions of years.

10.1 Timescale of cosmological events leading up to solar system formation

Based on precise observations of the cosmic background radiation, the age of the universe, that is, the time since the Big Bang explosion, is 13.7 billion years with an uncertainty of roughly 200 million years. This is not the only constraint on the age of the cosmos, however. The long-lived radioactive isotopes ^{238}U and ^{232}Th are both produced in the *r* process occurring in supernova explosions. Measurement of their abundances in stars with very few heavy elements relative to the Sun is possible using spectroscopy. Such stars are assumed to have formed not long after the cosmos itself, since they contain very few elements heavier than hydrogen and helium. When combined with models of the production of these isotopes by the *r* process, and with measurement of their abundances in meteorite samples, the time when production of elements heavier than helium began, in our Milky Way Galaxy, is constrained to be about 14 billion years ago. However, there is a large uncertainty of 3 billion years in either direction, in part because the Milky Way Galaxy may have been contaminated by material falling in from the intergalactic medium over billions of years but the rate of that contamination and its persistence are poorly known. Other age constraints from modeling ages of clusters of stars and cooling of white dwarfs narrow the time when stars – the factories of element production – first appeared to 12.5 billion years plus or minus one billion years. The two ages – one from the cosmic background radiation, the other from radioisotopes and stellar evolution models – are fully consistent with each other.

The beginning of element formation defines the earliest epoch of the evolution of our galaxy and others in the universe through the processing of hydrogen and helium in the nuclear furnaces inside stars. Two generations of stars likely were formed before the gas of our galaxy, now enriched in heavier elements and dust, brought into being our Sun and the planets of our solar system. Isotopic dating of meteorites using the rubidium–strontium system and others indicates that the most primitive rocks in the solar system are 4.56 billion years old, determined rather

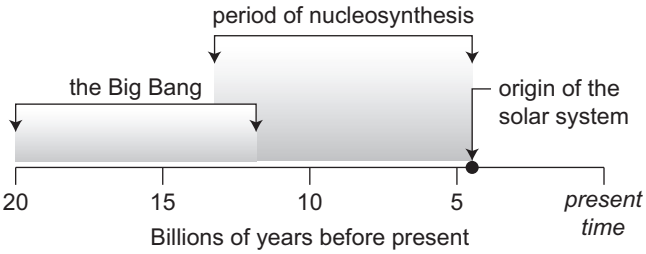


Figure 10.1 Timeline of cosmic events from the Big Bang through formation of the solar system.

precisely (Chapter 5). This is the time when solid matter was first assembled into what would become the solar system. At that epoch in galactic history, roughly 0.1% by number of the atoms in the galaxy were in the form of elements heavier than helium, and these played a key role in processes that led to planet formation.

Figure 10.1 sketches the cosmological timescale. What we seek to understand is how the material available in galactic clouds of gas and dust some 4.6 billion years ago evolved into a system of the Sun and the planets, and whether such a process is likely to be a common one. Twenty years ago, much of the discussion attached to this subject was speculation because of a lack of observations. Today, ample evidence in nearby star-forming regions exists of structures that appear to be the precursors of planetary systems. With new technologies applied to astronomical observations, we are able to identify and study the prenatal structures in stellar nurseries out of which planetary systems likely form.

10.2 Formation of stars and planets

10.2.1 Molecular clouds and star formation

Observations reveal regions of gas and dust dispersed among the hundreds of billions of stars that make up the spiral arms of our galaxy. The largest of these gaseous and dusty regions, *giant molecular clouds*, contain enough gas to make 100,000 stars each the mass of the Sun. The closest major molecular cloud, the Orion molecular cloud, is some 1,500 light-years from Earth. It is so large that it spans an area of the sky equal to 15 full Moons, but is not visible to the eye because it is so dim (Figure 10.2).

Most of the gas in molecular clouds is hydrogen and helium but, as noted above, one out of every thousand atoms in the molecules making up the gas and dust is an element heavier than helium. Virtually all of the atoms are combined into molecules, because of the low temperatures in the cloud and the relatively large number of atoms packed into every cubic centimeter

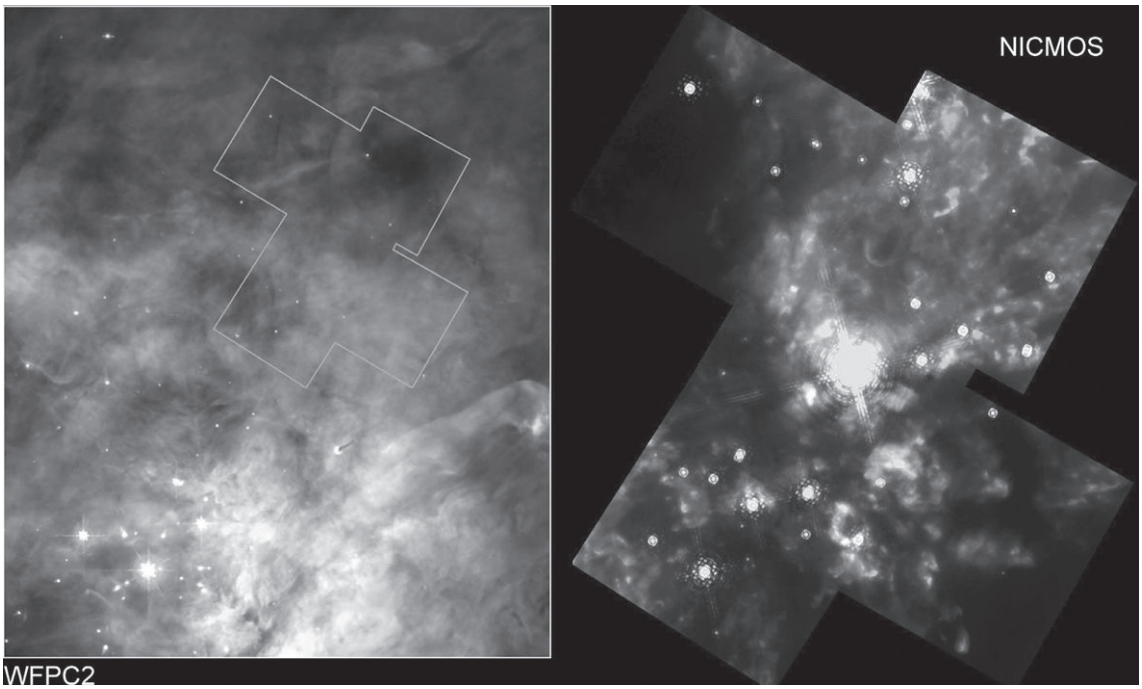


Figure 10.2 (Left) Sharpest image ever taken of the Orion Nebula, where star formation is occurring in a complex tapestry of environments of differing temperature and density some 1,300 to 1,500 light-years from Earth. In the bright central region of the image, called the trapezium because of the arrangement of stars there, ultraviolet light from massive stars is carving out a cavity in the nebula and possibly disrupting star formation there. Image taken by a team led by Massimo Robberto using the Advanced Camera for Surveys on the Hubble Space Telescope. (Right) Hubble near-infrared image of the boxed region reveals newly forming stars hidden by dust in the left-hand panel. NICMOS image by Rodger Thompson, Marcia Rieke, Glenn Schneider, Susan Stolovy (University of Arizona); Edwin Erickson (SETI Institute/Ames Research Center); David Axon (STScI); and NASA. WFPC2 image by C. Robert O'Dell, Shui Kwan Wong (Rice University), and NASA. For color version see plates section.

(i.e., high gas density) compared to other cosmic environments. In addition to molecular hydrogen (H_2), many different kinds of molecules occur with abundances that vary in complex ways from cloud to cloud and even within the same cloud. In the colder parts of molecular clouds many or most of the molecules are bound up in rocky and icy grains.

Determining the abundances of molecules in neighboring molecular clouds, some hundreds to thousands of light-years from the solar system, depends on the technique of spectroscopy (Chapter 3). Because temperatures in molecular clouds are low compared to those at the surfaces of stars, most of the photons emitted from the clouds are at long wavelengths, microwave parts of the spectrum. Where stars are forming, dust and gas falling into the nascent star may be heated to high temperatures, and light in the infrared and optical parts of the spectrum can be observed as well. Very precise microwave spectroscopy, such that the light is split into very fine wavelength bins so that spectral lines can be measured precisely, allows not only composition but also velocities of the gas to be determined. This in turn allows astronomers to map out regions of infall or collapse of gas and dust into nascent stars.

To get a sense of what a dense molecular cloud corresponds to in terms of terrestrial conditions, consider that the air in the room that you are occupying holds over 10^{19}, or ten million trillion, molecules of air, mostly nitrogen and oxygen, in every cubic centimeter. The average space between the stars, *interstellar* space, holds about 1 atom of hydrogen in each cubic centimeter of space; under these conditions, hydrogen is in the form of individual atoms rather than molecules of H_2. The densest clumps of dust and gas in a typical molecular cloud have 10^5 (100,000) atoms per cubic centimeter. Again, the density in clouds is determined from observing spectral lines of common molecules, such as CO (carbon monoxide), and tracing changes in the strength and shape of the line in different regions of a molecular cloud. Conditions – temperature, density, abundance of different molecules – vary widely between different parts of a given molecular cloud, ranging from cold tenuous portions grading into interstellar conditions all the way to dark, dense, localized clumps.

Most molecular clouds contain very bright but small areas of elevated temperature and strong energy emission. These glow in the infrared and their energy distribution (number of photons as a function of wavelength) is well simulated by computer models of stars surrounded by gas and dust. Such stars, called *T-Tauri* stars after the first one discovered, are very bright and are best explained as newly formed stars, or stars still forming by the processes described below.

10.2.2 The start of star formation

It appears, then, that molecular clouds are sites where stars form, so that portions of the Orion molecular cloud might be akin to that from which the Sun and the planets formed 4.5 billion years ago (Figure 10.2). Why do stars form in such clouds? There is plenty of hydrogen, helium, and heavier trace elements available to form the stars. The key to the stars' formation lies in the high-density regions of the cloud, which are gravitationally unstable. The dense dark clumps of molecular clouds are cold, and calculations show that the density of dust and gas is high enough that the mutual gravitational pull of the gas and dust should cause the material to come closer together, that is, to become denser. As the stuff becomes denser, the mutual gravitational pull becomes stronger and stronger, further increasing the density. This instability continues ad infinitum, the material falling into a common center and attracting more and more gas and dust.

How do such unstable clumps arise? Molecular clouds tend to inherit high levels of internal turbulence (random motions on scales larger than the space between atoms or molecules), and this causes the disk to both fragment and to create local clumps of denser material. Not all clumps can collapse further: the cloud of gas is threaded with magnetic fields, which attract charged particles in the gas and force them to move along the magnetic lines of force. The charged particles are a small but important fraction of the gas and, as they collide randomly with the neutral (uncharged) particles, they impart a pressure to the whole gas, a pressure caused by the force of the magnetic field on the charged minority in the gas.

This process prevents further collapse in some clumps, but not in the densest. Charged particles exist in the gas because neutral particles absorb high-energy light – ultraviolet (uv) photons, defined in Chapter 3 – from stars embedded in the cloud. These charged atoms, or ions, last only a certain time before they capture free electrons and become neutral again. Thick clumps of gas prevent the uv photons from traveling far, and so the thicker clumps of gas in the cloud have fewer charged atoms. The fewer the charged particles, the less pressure that is exerted on the gas by the magnetic field. Thus, the magnetic field is least effective at inhibiting the collapse of the densest cloud fragments. The cloud therefore is in a state of unstable equilibrium where, if a clump of gas forms of sufficient density, it will lose its ion population, lose its magnetic support, and begin collapsing to form a denser and denser core. This collapsing core is the beginning of the formation of a star or group of stars.

10.2.3 A star is born

As a core collapses in the molecular cloud, material falls deeper and deeper into the core's gravitational well, deepening the well. The molecules making up the gas and dust collide with increasing vigor toward the center of the core, converting uniform motion of collapse into heat. Temperatures at the center of the core become enormous – tens of millions of degrees – and pressures rise to billions of atmospheres according to computer simulations. (Astronomers can measure the brightness of such cores in molecular clouds such as that in the constellation Orion, which helps constrain these calculations.) Recall from Chapter 4 that these conditions are enough to initiate the fusion of hydrogen into helium, with release of energy. The energy generated creates a tremendous outward pressure in the core – the implosion of the gas has created an explosion at the center. A balance is achieved between the outward and inward pressures: too much expansion shuts off the fusion, reinitiating collapse, whereas too little expansion causes further implosion, a faster

fusion rate, and higher outward pressure. This newly balanced core of fusing hydrogen, surrounded by infalling gas and dust, is the picture that astronomers and physicists have developed of a newly formed star.

10.2.4 Figure skaters and astrophysicists: the formation of planets

Is there room in this picture for planets? Indeed, the formation of planets may be a natural consequence of the intrinsic spin or angular momentum of the gas. The entire Milky Way Galaxy consists of stars and gas moving in orbits about a common center. This circular motion is not completely uniform and, in particular, the gas in molecular clouds has eddies and turbulence that provide an intrinsic spin to the gas. A fundamental law of physics is that *momentum*, the product of velocity and mass of an object, is conserved; that is, it will not change unless a force acts upon it. This holds true for momentum associated with spinning motion, called *angular momentum*.

As a clump of gas collapses to form a core and then a newborn, or *proto-*, star, the gentle spin intrinsic to the extended tenuous gas becomes faster and faster as the clump becomes more compact. Why? To conserve angular momentum, the gas spins faster as it becomes more compact. The effect is just that of a figure skater: as the skater's arms contract she will spin faster even if she imparts no further force with her skates. (For this to work, her contact with the floor must involve little friction, hence the desirability of ice.) The collapsing core of a molecular cloud must shrink by a factor of 10^8 to become the size of a typical star like the Sun. Long before this size is reached, the spin rate of the gas becomes too large to allow continued infall to the center: the angular momentum forces the gas into an orbit around the protostar, along the spin direction. Thus a disk is formed within the collapsing gas, but if the angular momentum of the original clump is too high, it actually splits into two cores to form a binary star. This process is complicated: some of the gas, with little spin or angular momentum, falls right to the center. The rest is arranged according to angular momentum, with the gas having the highest angular momentum on the outer edge of the disk.

It is remarkable that most of the mass of our solar system is in the Sun and most of the angular momentum is in the planets. The disk out of which our solar system formed had to have possessed efficient mechanisms for moving mass to its center while retaining angular momentum in the dwindling disk material. Much of the extensive computer simulation work to understand the nature of disks from which planets form has focused on how enough angular momentum and mass could be transported in opposite directions (outward versus inward) during the limited lifetime of the disk. The lifetime itself is set by astronomical observations, which show that stars that are older than a few million years (based on spectral appearance and models) generally do not possess massive gas and dust disks.

Conceptually, it is possible to divide the evolution of a protoplanetary disk, or (for our Sun), "solar nebula," into four stages, as has been done by the Harvard astrophysicist A. G. W. Cameron. The rationale for such a division lies as much in conceptual convenience as it does in observations. It is likely that, if one could watch the evolution of such a disk, one would see the stages merge into each other and vary in their distinctiveness from one disk to another.

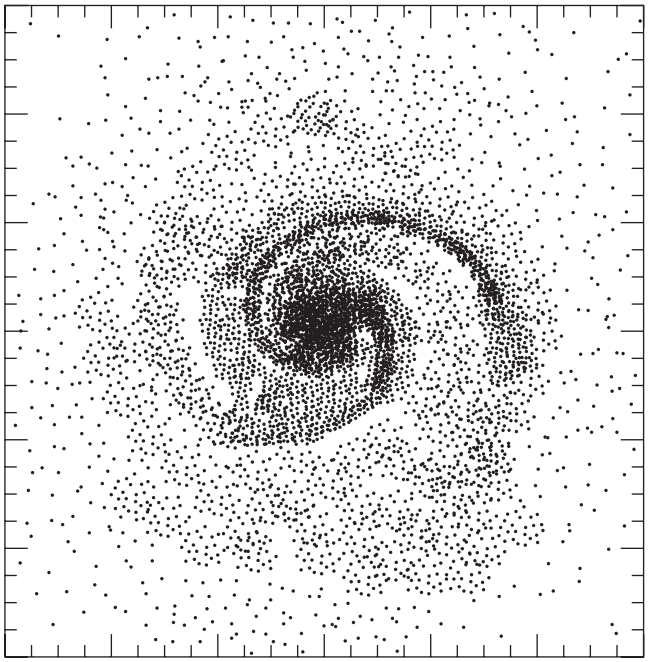


Figure 10.3 Spiral density wave pattern in a computer model of the protoplanetary disk, that is, the solar nebula, from which the solar system formed. The view is looking down on the face of the disk with the growing Sun (too small to be shown in this simulation) at the center of the figure. The disk is represented in the model by 8,000 discrete points; in reality the solar nebula was made up of countless more gas molecules and grains. The spiral pattern seen in the disk is reminiscent of the much larger scale structure seen in spiral galaxies. Simulation by A. Nelson, D. Arnett, W. Benz (University of Arizona), and F. Adams (University of Michigan).

The four stages are:

1. *Formation of the nebula*. The parent molecular cloud collapses to form a disk, perhaps because of the loss of magnetic support, as described above. The amount of material per square meter (the *surface density*) in the disk is increasing. This stage lasts perhaps a few hundred thousand years, very short compared to other astrophysical timescales.
2. *Dissipation in the nebula*. As material is added to the disk, some of it falls into the very center, forming the core of what will become the central star. The gas and dust in the disk begin to interact in three important ways. The heating of the disk sets up circulations of gas and dust, causing eddies that convert motion into heat and transfer angular momentum outward through the disk. Also, the gravitational force of material in the disk sets up waves in the gas, creating a pattern very similar to that seen in spiral galaxies (Figure 10.3). These waves act to create a force on the disk that causes further outward transport of angular momentum and heating. Finally, a small fraction of the gas is in the form of charged particles that are forced to move in a direction different from the bulk gas, because of the remaining presence of a magnetic field. All three of these processes – eddies, spiral waves, and magnetic effects – cause energy of rotation to be lost as heat,

Figure 10.4 Hubble Space Telescope image of a jet of material ejected from a disk of gas and dust surrounding a newly formed star. The star is hidden in the lower left portion of the image behind a disk of gas, dust and associated debris. The jet stretches outward trillions of kolometers from the start. This Wide Field and Planetary Camera-2 image courtesy of NASA and the Space Telescope Science Institute. For color version see plates section.

forcing more material to fall inward while shedding angular momentum to the outer extremities of the disk. The stage of most vigorous dissipation lasts perhaps 50,000 to 100,000 years. Evidence for it comes from disk systems, located in other star-forming regions, which suddenly brighten as seen from Earth; the best-studied example is a disk around the star FU Orionis in the Orion star-forming region.

3. *Terminal accumulation of the star*. Accumulation of more gas and dust has slowed dramatically. A wind of charged particles emanating from the star acts to erode the disk from the inside out; the present-day *solar wind* is the pale shadow of this primordial gale. Material also is ejected along the poles of the newly formed star in spectacular jets (Figure 10.4). Within the disk, the building blocks of planets – grains of rock and ice – are agglomerating together to form comet-sized bodies called *planetesimals*. In our own protoplanetary disk that became the solar system, the giant planets must have formed during this time, before the gas of the disk was blown away by the wind. This stage lasts several million years. Stars in such a phase are readily visible in molecular clouds because of the action of their winds; they are called T-Tauri stars after the best studied example of their class.
4. *Residual static nebula*. The central star has finished growing and is shining stably by virtue of hydrogen fusion. The vigorous wind that eroded away the nebula in stage (3) has largely but not completely abated and continues to drive off residual gas. Rocky planetesimals near the star agglomerate to form planets such as (in our solar system) the terrestrial planets. Observations of residual disks of dust around other stars, such as the star Beta Pictoris, whose disk was first imaged in 1984, suggest that this stage lasts from a few million to 30 million years.

Are the disks themselves, vastly smaller than the grand lanes and billows of the molecular cloud, observed? Until a decade ago, the answer was no. But a wide variety of techniques are used today to observe the disks of gas and dust around newly forming stars.

10.2.5 Disks around protostars: the source of planets?

The stages of star and disk formation outlined above can be observed indirectly or directly in the Orion and other neighboring molecular clouds. However, the act of planet formation in disks has never been observed. We have roughly 500 definitive examples of planets around normal stars, besides our own solar system; these exoplanets were detected beginning in 1995 using a variety of techniques described in section 10.5. The idea that Earth and the other planets formed from a disk of gas and dust is centuries old. The co-planarity and common orbital direction of the planets of our solar system led seventeenth century scientists to propose such a hypothesis. Beginning in the 1960s, study of the putative properties of the disk, called the solar nebula (meaning gas around the early Sun) has been based

on analysis of planetary atmospheres and primitive meteorites. Observations of disks in other star-forming regions, beginning in the 1980s, lent additional support to the notion that this is how planets form.

The source of planet formation in a disk is the turbulent motion of dust and gas. As material of different angular momentum sorts itself out according to distance, it collides with other material and generates heat. The collisions tend to cause material to fall ever inward until most of it ends up in the protosun. However, some of the dust sticks upon collision. The process of sticking, or accretion, can continue to ever larger sizes, from dust to pea gravel to golf balls to boulders.

The nature of the dust depends on position in the disk, and evolves with time. As accretion of material into the disk slows, the disk cools. In the inner disk, collisions of gas and dust are vigorous and heat the gas to hundreds or even thousands of degrees throughout the lifetime of the disk. Where the temperature is below 1,500 K, abundant rocky and metallic grains can survive: this region from 0.5 to 5 AU from the Sun is today the realm of the inner planets. Beyond about 5 AU, the gas was cold enough during much of the history of the disk to allow water ice to condense out and survive, and so, grains of both ice and rock were stable. It is in the outer solar system that we see bodies made of rock and ice – the moons of the giant planets.

Typically, material grows rapidly to bodies the size of our Moon or possibly Mars, after which growth slows dramatically because the spacing between these bodies has become large and collisions much less frequent. This phase of "oligarchic growth", so-called because the resulting bodies are roughly comparable in size to each other and few in number compared to the residual detritus of much smaller bodies, could be the end of the formation process. However, where giant planets are present, the orbits of the larger bodies are perturbed so that they are eccentric, allowing collisions and further growth to occur. This process of final agglomeration of large bodies on perturbed orbits can be simulated by a computer, and indicates that on a timescale of tens of millions of years, bodies the size of Earth can be formed (Figure 10.5).

Although this process of oligarchic growth followed by massive collisions explains the rocky terrestrial planets – Mercury, Venus, Earth, and Mars – it does not directly account for how the giant planets achieved their size.

The composition of Jupiter and Saturn differ from the Sun in being enriched in elements heavier than hydrogen and helium; some of this heavier material appears to be concentrated in cores at the centers of these planets. Uranus and Neptune are smaller objects that are a bit like Jupiter and Saturn but with most of the hydrogen and helium envelopes absent.

One explanation for the internal structures of the giant planets is that their formation started with the accretion of rock and ice, which produced a body large enough to gravitationally attract the gas of the solar nebula. As the gas concentrated near the growing planet, the gravitational field increased, drawing yet more gas, ice, and rock into the planet. Based on computer simulations, Jupiter and Saturn could have formed this way in a few million years. Uranus and Neptune may have taken longer to form, perhaps up to 10 million years longer based on recent computer simulations, and literally ran out of gas to make the envelopes as the solar nebula dissipated.

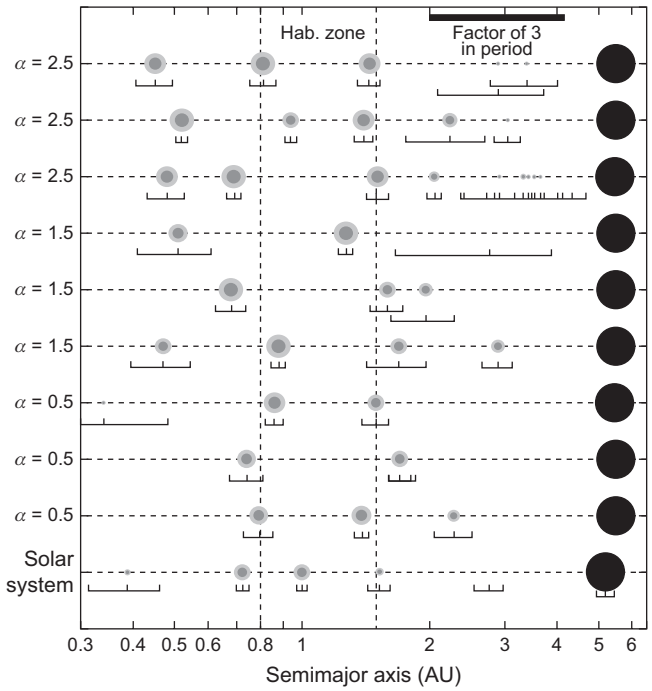


Figure 10.5 Nine computer simulations of the formation of planets with Jupiter present (large gray circle). The starting condition is a few hundred Moon to Mars-sized bodies distributed in different ways. For $\alpha = 2.5$ most of the mass is contained inward of 1 AU, for $\alpha = 0.5$ it is mostly beyond 2 AU, and $\alpha = 1.5$ is an intermediate case. The solar system shown for comparison. The size of each body corresponds to its relative physical size, but is not to scale on the x-axis. The dark circle in the center of each planet represents the size of its iron core. The eccentricity of the orbit of each body is shown beneath it, by its radial excursion over an orbit. Adapted from Raymond *et al.* (2005).

Alternatively, in very cold and quiescent disks, gas giants might have formed directly by collapse of the gas in the outer disk, but would then have acquired the same abundance of heavy elements as their parent stars. In some extrasolar giant planets, this seems to be the case, but not for others. Evidently both types of processes – collapse of gas onto a core or direct collapse of the disk gas – occur to form giant planets around different stars.

As the giant planets formed, they produced disks of gas and dust out of which their satellites, or moons, formed. The formation of Earth's Moon, which is not too much smaller than Earth, occurred a different way, and this is discussed later. The formation of Pluto and the other large members of the *Kuiper Belt*, the class of objects that were left over from planet formation, is less clear. But evidently collisions and growth occurred in that region during and after the formation of the giant planets. Comets, 10-km agglomerations of ice and rock, are incredibly numerous in orbits beyond Pluto, perhaps totaling to tens or hundreds of earth masses. They are the leftover detritus of planetary accretion propelled by encounters with the giant planets into far-flung orbits.

10.2.6 The end of planet formation

As the newly formed Sun reached a steady state between collapse and outward pressure from hydrogen fusion, its tremendous

energy generation produced not just a high luminosity of photons but a wind as well: a tenuous atmosphere of charged particles pushing outward from the solar atmosphere. We detect such a wind from the Sun today by spacecraft. Early in the Sun's history, the wind may have been enormous, profoundly affecting the entire solar system. Astronomers have observed in newly formed T-Tauri stars in other molecular clouds powerful winds that are driving the gas and dust away from the new stars. They also observe slightly older stars that have lost much or all of the surrounding gas and dust, and are no longer true, classical T-Tauri stars.

In the case of the Sun, this strong wind would have dispersed the solar nebula, stripped the early terrestrial planets of much or all of their original atmospheres, and driven very small grains of dust out of the solar system. With the gas and dust gone, accretion of new planets would have stopped because they were robbed of the raw material required.

The blowing-away of gas by wind and uv radiation from the newborn star extends beyond the circumstellar disk itself to the surrounding clumps of gas and dust in the molecular cloud. As star formation reaches a crescendo, and there is some evidence that it may be episodic, winds and uv radiation from multiple star systems erode significant portions of the gas and dust. A graphic example of this process comes from a Hubble Space Telescope image of a portion of a molecular cloud called the Eagle Nebula (Figure 10.6). Lanes of gas and dust are interspersed with clear areas, the whole illuminated by the light of newly formed stars that are driving the removal of the placental gas and dust.

Based on observations that constrain the lifetime of gas around other young stars, as well as clues from radioisotopic measurements as to when the Earth's core formed (Chapter 11), the formation of the solar system took of the order 10 million years. By this time, the Sun had settled into a steady state and the planets of the solar system, including Earth, were in place. Although the formation of the terrestrial planets, including Earth, took up to 100 million years, the giant planets must have formed more quickly, in less than 10 million years, so that the gas of the solar nebula was still present.

Figure 10.6 Hubble Space Telescope image of the Eagle Nebula, showing clearing of the gas and dust by ultraviolet light from newly formed stars, which are illuminating the scene. The image was taken by Jeff Hester and Paul Scowen of Arizona State University using the Wide Field and Planetary Camera 2 at visible wavelengths. Courtesy of NASA and the Space Telescope Science Institute. See color version in plates section.

10.3 Primitive material present in the solar system today

The planets have evolved through time after formation, their composition being altered by *outgassing* and *chemical differentiation*. Details of these processes are presented for Earth and the other planets in Chapter 11. Because no planetary body, even the larger moons, has escaped this evolution, what record do we have of the original composition of solid material throughout the solar system? Some meteorites have been known for many years to have a composition that mimics that of the Sun over certain classes of elements, and these are almost certainly unevolved remains from the solar system's formation. Tiny bits of dust that make their way into Earth's atmosphere – *interplanetary dust particles*, or IDPs – also appear to be unaltered samples of rock-forming materials. If we want to examine primitive *ices* (water, ammonia, carbon dioxide, methane, nitrogen, carbon monoxide, among others) we must go farther out in the solar system to Kuiper Belt objects and comets, discussed further in section 10.4.2.

10.3.1 Remnants of the beginning: meteorites

An important record of the earliest time in the history of the planets, when large solid bodies first were present in the solar nebula, is in meteorites – rocks that fall to Earth from the sky. The origin of meteorites lies in larger *parent* bodies from which fragments have been blown off by impacts. The wide range of meteorite types, from nearly pure iron and nickel to those with largely silicate (rocky) and organic (carbon-bearing) composition, require that they come from at least several dozen distinct parent bodies. Most of these are likely to be asteroids either in the main asteroid belt or on other orbits in the solar system. About a dozen meteorites each are recognized to be from the Moon and Mars.

Some of the meteorites have abundances of elements over a certain range of volatility that are very similar to those in the Sun. (Recall that volatility is simply a measure of the tendency of a material to vaporize into a gas; water is more volatile than rock.) Elements that have an affinity for incorporating in rocky matter are generally all present. Elements such as carbon generally are depleted, but do exist in *organic* (carbon-bearing) phases of the meteorite. The more volatile elements apparently did not condense into the grains from which meteorites formed, but are present in significantly depleted amounts trapped in the rocky or organic phases of the primitive carbonacous chondrite meteorites (Chapter 5).

The general consensus is that the chondrites could have come from parent bodies that formed relatively close to the Sun – the asteroid belt, for example. Nonetheless, they might bear a strong resemblance to the rocky component of primitive bodies from farther out in the solar nebula, that is, comets. Differences between the chondrites and rocky components of comets would provide very strong constraints on the distinct processes that operated in the hot inner part of the solar nebula versus the cold outer part. These exciting comparisons await detailed examination of a comet nucleus.

The carbonaceous chondrites may have survived relatively unchanged from the period of planet accretion. Their texture, differences in elemental abundances from the Sun, and patterns of isotopic abundances tell us much about the solar nebula. In particular, the gas in the region where the terrestrial planets formed must have been hot enough, very early in the solar nebula's history, for most elements to have been in the gaseous phase, followed by condensation to form the primitive meteorites as temperatures in the nebula dropped. This may be a very different situation from the region of the outer planets, where the rocky component of solid material was never vaporized, but instead fell into the nebula as largely intact grains. (The water ice, on the other hand, likely evaporated and then recondensed, at least in the region corresponding to the orbits of Jupiter and Saturn.)

In addition to the primitive matrix material, chondritic meteorites contain tiny, nearly spherical, nodules called *chondrules*, which appear to be bits of the rock that melted, resolidified, but did not change significantly in chemical composition. The origin of the chondrules is a mystery, but seems to imply that there were sudden episodes of heating in the solar nebula, perhaps caused by electrical discharges akin to terrestrial lightning, or by bursts of energy associated with a strong magnetic field in the nebula. They illustrate the complex nature of the processes from which the planets formed.

10.3.2 Comets and Kuiper Belt objects

Beyond Neptune's orbit lie other remains of solar system formation, bodies ranging in size from tiny grains to minor planets hundreds of kilometers across (Figure 10.7). Some objects currently residing in the Kuiper Belt very likely have been there since the beginning of the solar system, while others may have been kicked there by the gravitational effects of the giant planets. The Kuiper Belt objects are likely to be a mixture of rock and ice, with very volatile species such as methane, nitrogen, and carbon monoxide still present as ices themselves, or trapped in void spaces in the water ice. The very low temperatures (30 to 50 K) in the Kuiper Belt ensure that evolution of these bodies is slow. So, modified perhaps slightly by collisions and by internal heat sources for the larger bodies, Kuiper Belt objects should preserve a record of the ice and rock composition of the original grains of the outer solar nebula, those that formed the moons and cores of the giant planets. Meteorites were part of parent bodies that formed too close to the Sun to have retained the volatile components as ices, and so, Kuiper Belt objects and comets provide a distinct record.

A visit to a Kuiper Belt object will happen in 2015 when NASA's *New Horizons* spacecraft flys by Pluto and its moons. Pluto is the second largest Kuiper Belt object after Eris; Pluto's orbit is very similar to a number of recently discovered smaller objects in the region and occupies a part of the Kuiper Belt that is not disturbed by Neptune. Until the first spacecraft flyby, these objects are so faint that chemical examination of their surfaces, using spectroscopy to break the light into component wavelengths, requires the largest Earth-based telescopes. Recent successes in identifying surface ices on Kuiper Belt objects, besides Pluto and it largest moon Charon, represent remarkable progress in astronomical spectroscopy.

Comets are more readily examined, and they too are icy bodies from the farthest reaches of the solar system. However, their detectability stems from their very noncircular orbits that bring them close to the Sun, vaporizing ices from the outer layers and lofting dust, that allows examination remotely from Earth. Material from only one comet, Halley, has been analyzed directly by spacecraft. Although comets contain a valuable record of the primitive composition of ices from the outer solar system, their very noncircular orbits that bring them close to Earth also make it difficult to understand where they originally formed. Although these orbits seem to have their origin in a region called the *Oort Cloud*, extending to 100,000 AU from the Sun (3,000 times the Pluto–Sun distance), it may be that comets did not form at such distances. More likely is that the Oort Cloud comets were formed closer in, where the giant planets reside today, and then were flung outward in gravitational close encounters with the giant planets (Figure 10.8). Also, certain classes of comet orbits, such as those possessed by the *Jupiter family short-period* comets, could not have evolved from the Oort Cloud, but instead these objects are likely to be Kuiper Belt bodies perturbed out of their original orbits by collisions or repeated gravitational perturbations by the giant planets.

Thus, the record of composition contained in comets is difficult to relate directly to a particular region of the early solar system, except perhaps in the case of the Jupiter family short-period comets. Nonetheless, the accessibility of comets makes them interesting. The European Space Agency plans a mission, Rosetta, to rendezvous with and land on a comet in 2014. An American robotic probe, *Stardust*, flew through the coma (dust and gas cloud) of a comet and collected dusty material, which was returned to Earth in 2006. Another US spacecraft, *Deep Impact*, in 2005 launched a cannonball to blast through the icy crust of a comet nucleus to obtain spectra of fresh ices during the flyby of the main spacecraft.

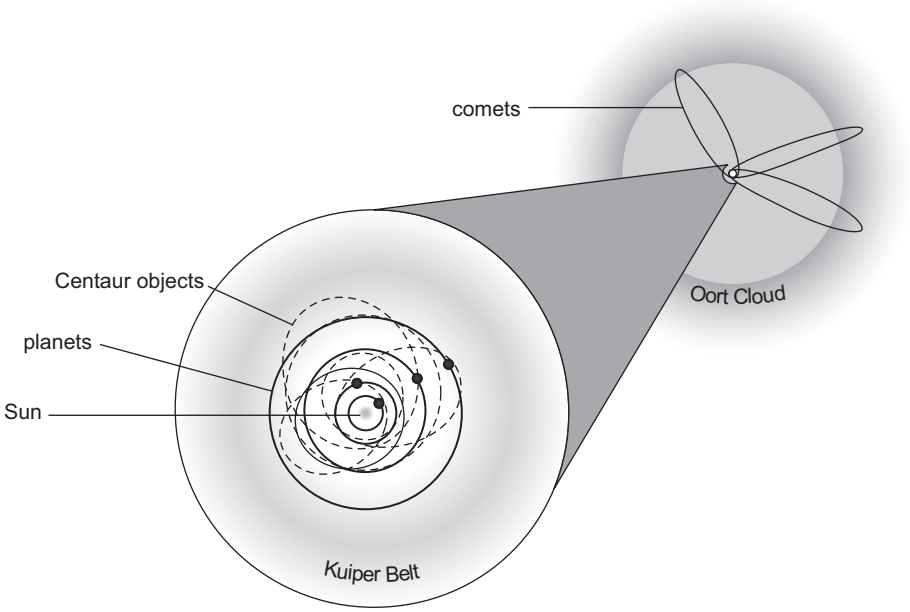


Figure 10.7 Sketch of the outer solar system, showing the location of the Kuiper Belt as the shaded region. The solid orbits are (from inside outward) Jupiter, Saturn, Uranus, and Neptune. Pluto's orbit, not shown, places that planet in a stable part of the Kuiper Belt. The dashed lines show orbits of Centaur objects, which probably have been perturbed inward from the Kuiper Belt by Neptune's gravity as well as, perhaps, through collisions in the Belt itself. On the upper right is a diagram showing the much larger Oort Cloud and a few sample long-period comet orbits. The Oort Cloud is so much larger than the Kuiper Belt that the latter can hardly be seen on the scale of the former.

10.3.3 Interplanetary dust particles

Interplanetary dust particles (IDPs) are microscopic bits of cosmic dust that enter Earth's atmosphere and fall slowly toward the ground, or are collected from Earth orbit. Their fluffy nature (Figure 10.9) and very small size ensure a gentle descent through Earth's atmosphere, and they can be collected by airplanes with appropriate sampling tools. The origin of IDPs is not certain, but their composition resembles the primitive carbonaceous chondrites. It has been suggested that they are derived from the dust (non-ice) component of passing comets. Unfortunately, IDPs are too close to the Sun for ice to be stable, and hence an important clue linking them to comets is missing. More detailed information on the silicate and organic components of comets is required to make this linkage.

10.4 The search for other planetary systems

The problem of identifying planets around other stars is one of enormous challenge. Jupiter's brightness, at optical wavelengths to which our eyes are sensitive, is one-billionth (10^{-9}) that of the Sun. At infrared wavelengths, where cooler objects tend to radiate (as described in Chapter 3), Jupiter is still a mere 10^{-5} as bright as the Sun. Imagine looking at our solar system from a great distance, many light-years away. Any normal telescope system will see the Sun, the central star, but even Jupiter is lost in the glare caused by light scattered across imperfections in the telescope mirrors and across the telescope structure.

To overcome these problems, astronomers pursue several types of approaches to detect planets around other stars. *Indirect techniques* rely on observing the effect of the planet on the motion or appearance of the parent star. *Adaptive optics* and *interferometric imaging* are both *direct* techniques to overcome telescope glare and the smearing effects of our own atmosphere (through which Earth-bound astronomers must look) to achieve images of planets orbiting another star.

10.4.1 Indirect techniques

Indirect techniques illustrated in Figure 10.10 involve watching the position of the star oscillate in the sky, caused by the gravitational effect of a planetary companion. Motion back and forth can be seen using precision position measurement determinations referred to as *astrometry*. Star wobble toward or away from the observer can be identified by looking at one or more spectral lines from the star that reveal the *radial velocity* through the Doppler effect described in Chapter 2. As the star moves toward the observer, the line is *blue shifted*; as the star moves away, the line is *red shifted*. The mass of the planetary companion can be determined from the magnitude of these effects.

Another indirect technique works if the orbit of the planet around the star is close to the line of sight to Earth. Then, as the planet passes in front of ("transits") its parent star, it blocks some of the light and causes a partial eclipsing of the starlight. The technique has been used from the ground, but the low probability that any given planetary system will be suitably aligned for transits makes a space-based system capable of viewing a large part of the sky attractive. Two such missions, Corot from the French and European Space Agencies, and Kepler from the US Space Agency NASA, are currently accumulating statistics from transits of the occurrence of Earth- and near-Earth-sized planets

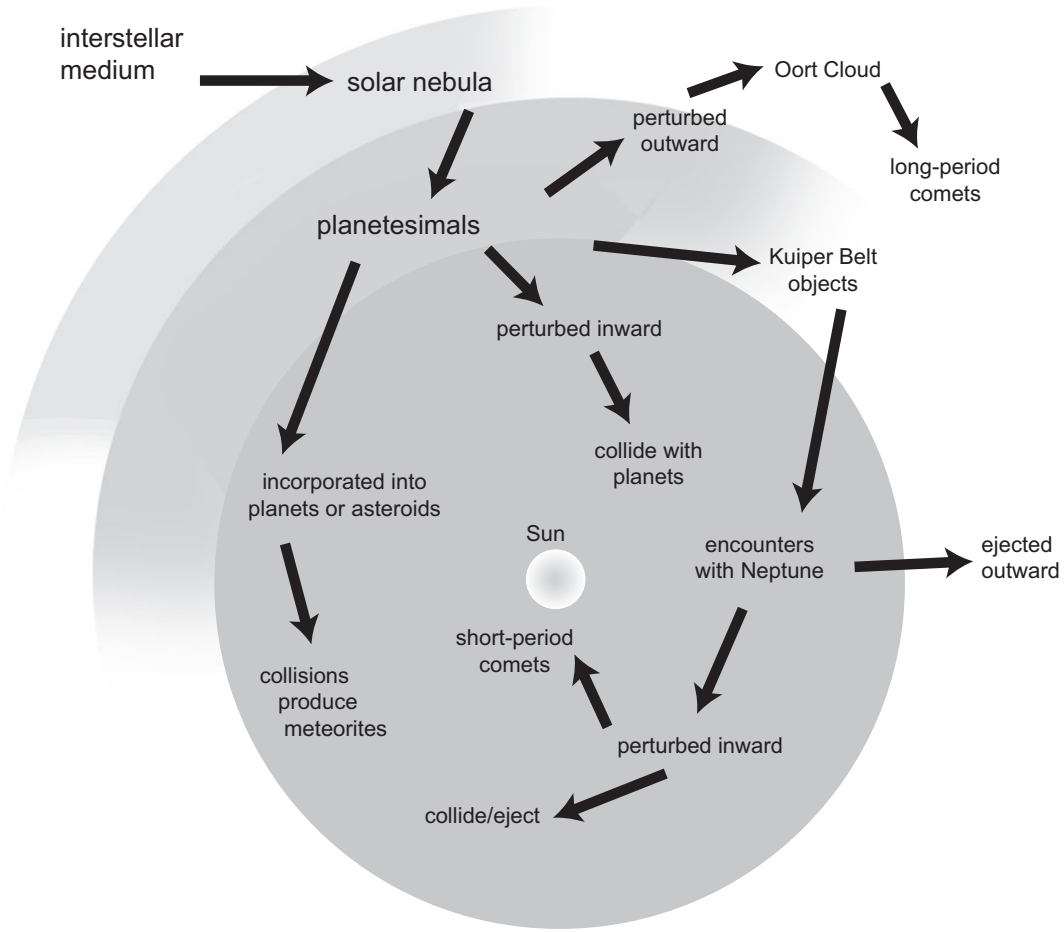


Figure 10.8 Possible history of small bodies in the solar system, from the original molecular cloud of the interstellar medium in which the solar system was born, through the solar nebula phase to the accumulation of planetesimals into larger bodies. Some planetesimals became part of the growing planets and asteroids; later collisions among asteroids produced fragments, some of which reach Earth as meteorites. As giant planets formed and their gravity increased, orbits of remnant planetesimals were increasingly perturbed; some planetesimals were ejected to the Oort Cloud, others inward to collide with the terrestrial planets. The Oort Cloud became the source of the long-period comets. Remnant planetesimals just beyond Neptune constitute the Kuiper Belt, and some of these have survived to the present day. Others, perturbed principally by Neptune's gravity, were either ejected outward or shunted inward to form Centaur objects. These either collide with the giant planets or have their orbits further altered to become short-period comets. Based on a scheme by Cruikshank (1997).

around other stars. It appears that such planets are common – perhaps 10% of stars like the Sun have such planets.

Microlensing detects planets through an entirely different effect. As seen from Earth, if a star passes in front of a more distant, background star, the light from the background star is temporarily enhanced by the bending of light rays around the nearer star, in accordance with the general theory of relativity, which predicts that gravitational fields bend light rays. If a planet is present in the right position around the nearer star, it produces a further brightening, which is distinguishable from that of its parent star. Though such microlensing events are rare, a modest-size telescope in space can automatically scan many hundreds of thousands of stars to catch those rare signatures of the focusing of light by a passing interloper and its planet.

Indirect techniques have yielded a plethora of important results. Since 1995, over 500 planets, from the mass of Jupiter down to just a few times the mass of the Earth, have been discovered around other stars by these techniques. Transiting planets with radii down to that of Neptune have been found by the *Corot* and *Kepler* spacecraft (Figure 10.11). For those planets seen both in transit and via the Doppler shift, both radius and mass are known so that we also know their density. It has been possible to infer that giant planets like Jupiter do exist around other stars, but they have a variety of different properties including composition, weather, etc. The atmospheric properties of some of these bodies have been discovered by spectra taken during transits of the planets behind or in front of their parent stars. And around the smallest "M-dwarf" stars, where detection is easier, planets only a few times the mass of the Earth have been seen.

In 1992, radial velocity techniques were employed in a very different fashion to detect planets around a pulsar. A pulsar is the ultradense neutron star core of an exploded star, one that has finished the chain of fusion reactions described in Chapter 3. Most such neutron stars have very strong magnetic fields, which result in charged particles streaming along the magnetic poles

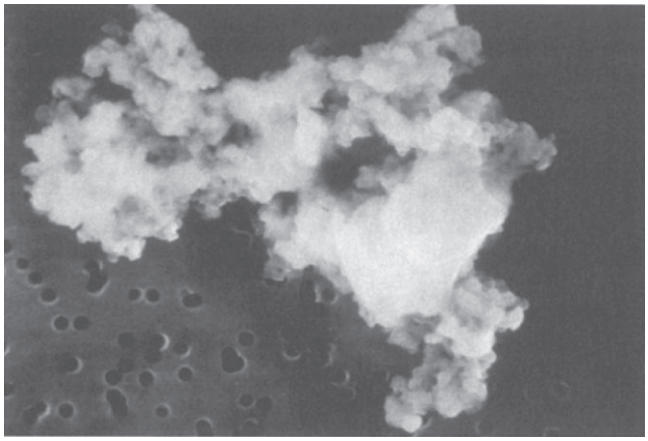


Figure 10.9 Photograph, using an electron microscope, of an interplanetary dust particle, roughly 10 microns across. The dark holes in the background (used to help mount the particle) are 0.4 microns across. Image courtesy of Professor Don Brownlee, Washington University.

of the star, creating a beacon that can be detected at radio wavelengths. Using the Arecibo radio telescope in Puerto Rico to measure the Doppler shift to progressively shorter (bluer) and then longer (redder) wavelengths of radio energy, National Radio Astronomy Observatories (NRAO) astronomer D. Frail and team were able to infer the presence of at least two and, from 1994 observations, possibly as many as four, planets orbiting the pulsar PSR1257+12. These planets range in mass from several times that of Earth to a mass as small as that of Earth's Moon.

How planets could have survived the pulsar-creating explosion of the original star is a mystery. One idea holds that

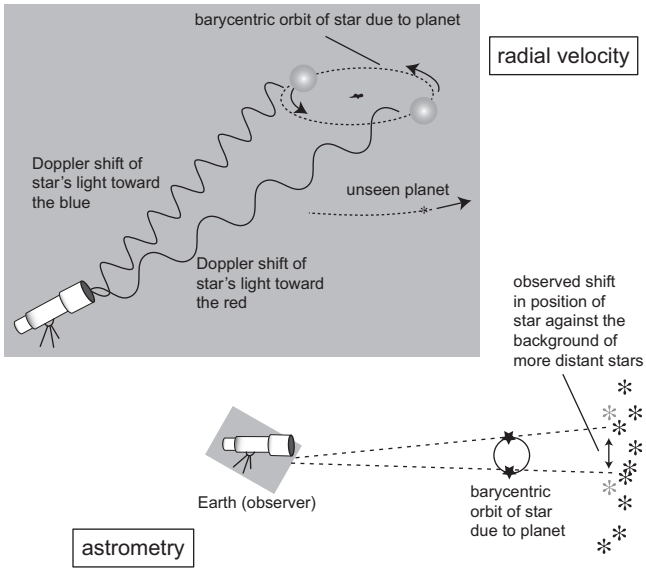


Figure 10.10 Examples of indirect techniques for detecting planets. Shown in both sketches is a star and its companion planet. The planet forces the star to be nonstationary, that is to orbit the common center of mass of the system. The observer on Earth is symbolized by the telescope, which is extremely far from the star and its companion planet. In the radial velocity technique, the distant observer on Earth sees the component of the star's motion (actually its velocity, *V*) directly toward or away from the Earth via Doppler shift. In the astrometric method, the star's slight shift side to side in the sky, due to its orbital motion, is detected on Earth. In both cases, the planet itself is lost in the glare of the parent star, and is detected only by its gravitational influence on the star. Adapted from a drawing by NASA.

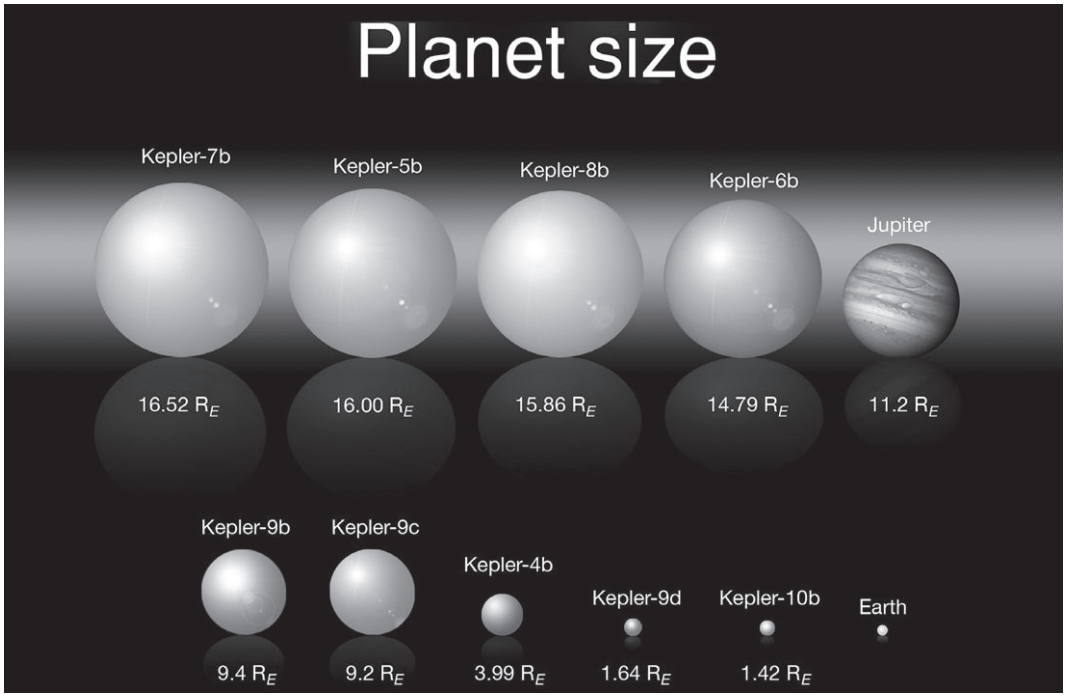


Figure 10.11 Planets found in the early part of the Kepler space mission using the transit technique. Sizes of the planets are shown compared to Jupiter and the Earth; R_E is the radius of the Earth. From a color figure courtesy NASA-Ames Research Center, Wendy Stenzel.

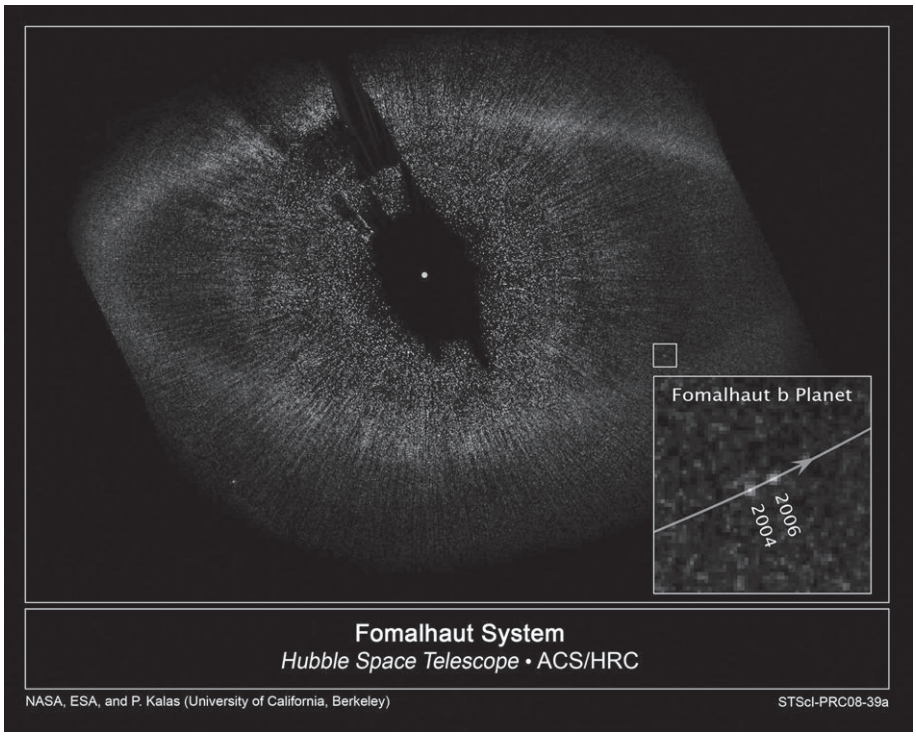


Figure 10.12 A disk of dust is seen around the star Fomalhaut at optical wavelengths, using the coronagraph onboard the Hubble Space Telescope to dim the light of the parent star. The inset is a composite image showing the location of a planet orbiting the star, seen in 2004 and 2006 relative to Fomalhaut. By looking at two succeeding years the motion of the planet can be detected, and indeed it seems to be moving in an orbit nested within the dust belt. From Kalas *et al.* (2008). See color version in plates section.

the planets did not exist prior to the supernova explosion but were instead created from debris of the explosion in a process mimicking planet formation around very young stars. The presence of these planets suggests that such formation processes can occur in many different kinds of environments around stars.

10.4.2 Direct techniques

There are two distinct approaches to suppressing the light from a star to a sufficient extent to see a planet orbiting around it: coronagraphy and interferometry. An internal coronagraph blocks the starlight using optical elements within a telescope, while an external-occulter coronagraph blocks the starlight with a separate large starshade positioned in front of the telescope, usually many tens of thousands of kilometers away. The chief advantage of internal coronagraphs is their simplicity in pointing and centering the coronagraph on the central star. However, there is a practical limit to the size of the telescope that can be used, because it must be launched into space. Coronagraphic observations have already yielded images of widely separated planets and, in one case, of an intervening disk (Figure 10.12).

The appeal of external coronagraphs, which have been studied for many years, is their potential to circumvent many of the light suppression problems faced by internal coronagraphs by instead blocking the stellar light with a free-flying starshade. The main drawback of the external-occulter approach lies in its operational complexity relative to a single spacecraft – two vehicles must perform properly for this technique to work and source targeting requires aligning the two spacecraft.

At long infrared wavelengths where the planet–star contrast shrinks by several orders of magnitude, coronagraphs would become huge and unwieldy. Instead, interferometery is the favored approach. An infrared interferometer consists in its simplest form of two telescopes joined on a structure, or mounted on separate satellites that maintain a controlled distance by precision formation flying. The starlight is suppressed by arranging the light beams coming into the two telescopes to be combined so that at the center of the image, they destructively interfere with each other and cancel the light of the glaring star. Light that is off center, such as from a planet displaced from the star, is not destructively cancelled and so has a much higher contrast than were the system observed with a single telescope alone.

10.5 Summary of planet formation

Twenty years ago our solar system was known as the singular example of the end state of the process of planet formation, with no observable examples of such systems actually in formation. Today, evidence exists that star-forming regions have abundant disks of gas and dust, and at least some of these likely evolve into planetary systems. We know of 500 planets orbiting stars other than the Sun, and although a system

like our own is difficult to detect, little by little systems resembling our own are being found. Terrestrial planets like Earth are very difficult to detect around stars like the Sun (they are much easier to see around cool red dwarfs), but the technology anticipated to be available within two decades will be able to do so. Indirect detection techniques have already indicated strongly that planets the size of the Earth should be relatively common. Here at the start of the second decade of the new millennium, we are close to having the answer to the question: is our solar system a unique or rare peculiarity of the evolution of stars and galaxies, or are we one of many such systems in the heavens?

Summary

Our Sun is a third-generation star in a Universe that is 13.7 billion years old; within a few hundred million years stars began generating heavy elements from the hydrogen and helium that was the primordial material formed at the Big Bang. The Sun itself is 4.5 billion years old, a product of debris contaminated with heavy elements produced in the billions of years preceding its formation. Stars form in clouds of gas and dust of varying size, composed mostly of hydrogen and helium but seeded with heavier elements from supernova explosions. Turbulence within the clouds, and possibly also compression from supernova explosions of the most massive (and short-lived) stars near the cloud edges leads to the formation of denser clumps of material. Such clumps should be gravitationally unstable, in the sense that their own gravitational pull should make them denser and denser until they literally implode. However, magnetic fields threading the clouds partially support the clumps against collapse, at least temporarily. As clumps collapse, the conservation of angular momentum amplifies the spin of the material, in some cases forming disks of material around the central mass, which eventually becomes hot enough to ignite fusion and become a star. The disk itself is the site, in some cases, of planet formation, a process that may occur in several episodes while the gas is present and then when it is dispersed by winds and ultraviolet radiation from the central star. The giant planets, rich in gas, must form before rocky planets like the Earth. The process of planet formation is recorded, however imperfectly, in the chemical composition of planets, moons, and small bodies, as well as in the arrangement of their orbits. Over 500 planets around other stars have been discovered by a variety of techniques, and these planets range in size from above that of Jupiter to several times the mass of the Earth. Indications from current planet searches are that perhaps 10% of stars like the Sun possess planets within a few times the size of the Earth.

Questions

1. Given that giant planets have been discovered very close to a handful of stars (much closer than Jupiter is to our Sun), does the picture of planet formation presented here require revision? How would you revise it?
2. What kinds of measurements would you make to determine whether IDPs come from comets or asteroids, and if the latter, what particular type of asteroid?
3. From the conservation of angular momentum, calculate the factor by which the rotation rate of material with a given angular momentum increases when a disk is formed from a molecular cloud clump (hint: you will need to look up the sizes of the clump and typical protoplanetary disks).
4. Go to the site http://exoplanet.eu/, the Exoplanet Encyclopedia, and peruse the list of planets. Pick a planet that particularly interests you, and write a paragraph on how it was discovered, known properties, resemblances to planets in our own solar system, etc.

General reading

Reipurth, B., Jewitt, D., and Keil, K. 2007. *Protostars and Planets V*. University of Arizona Press, Tucson.

Seager, S. 2010. *Exoplanets*. University of Arizona Press, Tucson.

References

Angel, J. R. P. and Woolf, N. J. 1996. Searching for life on other planets. *Scientific American* **274**(4), 60–6.

Beckwith, S. V. W. and Sargent, A. I. 1993. The occurrence and properties of disks around young stars. In *Protostars and Planets III* (E. H. Levy and J. I. Lunine, eds). University of Arizona Press, Tucson, pp. 521–41.

Bergin, E. A. and Tafalla, M. 2007. Cold dark clouds: the initial conditions for star formation. *Annual Reviews of Astronomy and Astrophysics* **45**, 339–96.

Bond, J. C., Lauretta, D. S., and O'Brien, D. P. 2010. Making the Earth: combining dynamics and chemistry in the solar system. *Icarus* **205**, 321–37.

Cruikshank, D. P. 1997. Organic matter in the outer solar system: from the meteorites to the Kuiper Belt. In *From Stardust to Planetesimals*, ASP Conference Series Vol. 122 (Y. J. Pendleton and A. G. G. M. Tielens, eds). Astronomical Society of the Pacific, San Francisco, pp. 315–33.

Dauphas, N. 2005. The U/Th production ratio and the age of the Milky Way from meteorites and galactic halo stars. *Nature* **435**, 1203–5.

Kalas, P., Graham, J. R., Chiang, E. *et al*. 2008. Optical images of an exosolar planet 25 light-years from Earth. *Science* **322**, 1345–8.

Mason, S. F. 1991. *Chemical Evolution*. Clarendon Press, Oxford.

McKee, C. F. and Ostriker, E. C. 2007. Theory of star formation. *Annual Reviews of Astronomy and Astrophysics* **45**, 565–687.

Raymond, S. N., Quinn, T., and Lunine, J. I. 2005. Terrestrial planet formation in disks with varying surface density profiles. *Astrophysics Journal* **632**, 670–6.

11

The Hadean Earth

Introduction

The period from the formation of Earth, some 4.56 billion years ago, to the time when the oldest rocks still in existence today were formed, roughly 3.8 billion to 4.0 billion years ago, is called both the *Hadean* era and *Priscoan* eon of Earth. The term Hadean, referring to the classical Greek version of hell, is well chosen, because all evidence that we have is that the Hadean Earth was very hot and extremely active, with widespread volcanism and frequent impacts of debris left over from planetary formation. This time encompasses the assemblage of Earth from the smaller *planetesimals*, dramatic internal rearrangements such as core formation, the creation of the ocean and earliest atmosphere, and the origin of Earth's Moon. Forces that acted on Earth were essentially the same as those acting on Mars and Venus, and a traveler visiting Earth would have seen little to distinguish it from the two neighboring terrestrial planets.

Each planet initially had a molten, or nearly molten, silicate surface, followed by cooling and establishment of a solid crust. Each had an atmosphere dominated by carbon dioxide (CO_2), with little free molecular oxygen (O_2). Evidence exists that each planet had liquid water on its surface during a portion of the Hadean era. Most important, no sign of life could be seen on any of these three planets – conditions were too severe and variable to allow life-forms to survive except near the end of the Hadean on Earth, and perhaps at about the same time on Mars.

The close of the Hadean saw a fall off in impact rates as the solar system's planets swept up much of the remaining debris, and on Earth the stabilization of a liquid water ocean and the build up of chemically processed, buoyant crustal fragments that stood above the level of the sea: protocontinents. It was during this time, perhaps 4 billion years ago, that life began to survive continuously on Earth. Life also might have gained a foothold on Mars, and even briefly Venus, as well, but beyond the Hadean, Earth's evolution diverged from that of its planetary neighbors in terms of its atmosphere, surface, and growing abundance of life.

In this chapter we focus on key events and processes on the Hadean Earth, and also take a close look at the chemistry of materials that build terrestrial planets. This *geochemistry* will be an important part of the story of interaction between Earth's crust, atmosphere, and life in evolving a habitable world.

11.1 Bulk composition of the planets

We can measure the masses of the planets using the periods and distances of their natural satellites (Kepler's laws, described in Chapter 3) or (for moonless Mercury and Venus) via observing their small gravitational effects on the slightly elliptical orbits of neighboring planets or on flyby spacecraft. Knowing their distances from us, we can measure the size of each of the planets using powerful telescopes and resolving the disk of the planet. Given the mass and the size, it is possible to determine the density of each planet, that is, the mass divided by the volume. The density for the terrestrial planets is given in Figure 11.1.

The bulk density of a planet is determined both by the material out of which it is made and the amount of compression of that material caused by the planet's own gravitational field. Because different materials compress to differing extents (for example, a soft pillow versus a slab of rock), the composition and compression are coupled. The more massive the planet, furthermore, the more compression. Also listed in Figure 11.1 are the uncompressed densities, that is, the densities of the planetary materials in the absence of self-compression. We can display these as single numbers only for the solid planets;

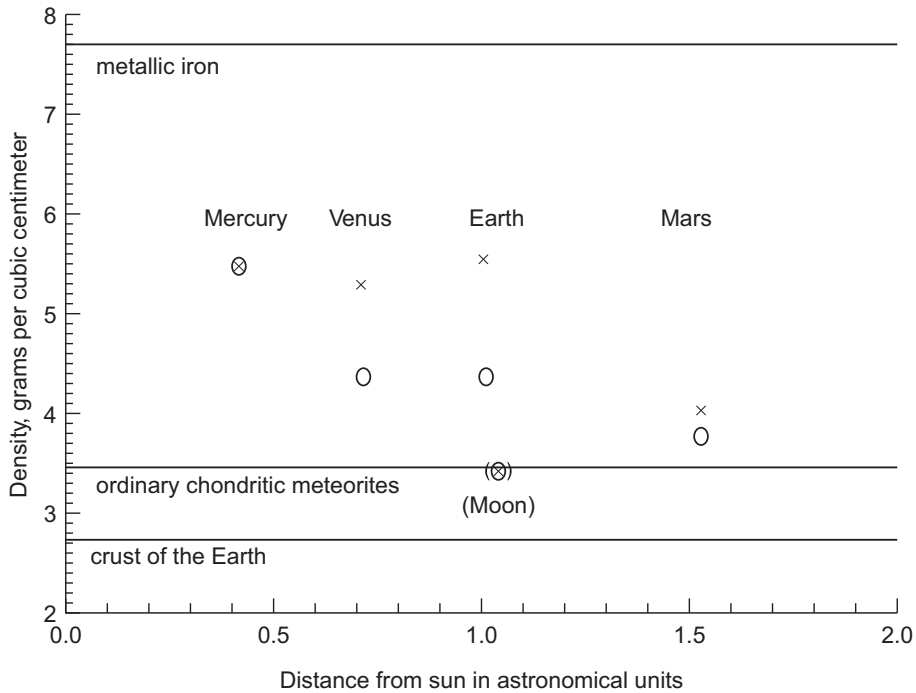


Figure 11.1 Densities of terrestrial planets and candidate mineral components. The planets are plotted as a function of their distance from the Sun. The symbol × indicates the measured density; the symbol ○ refers to the uncompressed density of the planet when the effect of gravitational compression of the material is removed. The compressed and uncompressed densities for the Moon are shown in parentheses, to distinguish them from Earth's. The density of Earth's crust also is shown; for comparison, the density of liquid water under low pressure is 1 gram per cubic centimeter.

for the giant planets, most or much of the uncompressed material would be an ideal gas for which the density depends on the pressure and temperature under which the gas is contained.

Using the uncompressed densities, the abundances of the elements shown in Figure 3.8 of Chapter 3, and some chemical knowledge of how these elements tend to combine in the interiors of planets, we can infer the most abundant constituents in each of the planets. Obviously, this is not a simple deductive exercise because the uncompressed densities given in Figure 11.1 were computed on the basis of some assumed composition. Instead, it is an *iterative* exercise, wherein one eliminates certain materials for certain planets because they do not produce the right compressed density. We will not go through the details of the exercise, but instead summarize the results for the solid planets, and then for the giant planets.

11.1.1 Solid planets

The designation solid planets includes the terrestrial planets (Mercury, Venus, Earth, Mars, and the Moon), Pluto, and the larger moons of the outer solar system. The terrestrial planets all have uncompressed densities in the range of 3 to 6 grams per cubic centimeter (g/cm^3); by comparison, the density of liquid water at normal conditions is 1 g/cm^3. Chondritic meteorites, composed to a large extent of minerals containing silicon, magnesium, and oxygen, have a density in the 3- to 4-g/cm^3 range. The meteorites provide us with clues as to the nature of planet-building materials in the right density range. In terms of cosmic abundance, the rock-forming elements silicon and magnesium are less abundant than elemental oxygen. The silicon and magnesium combine with abundant oxygen to form minerals such as enstatite ($MgSiO_3$), forsterite (Mg_2SiO_4), and other "rocky" compounds.

Densities of the silicate minerals are too low to account fully for the uncompressed density of all but Earth's Moon and perhaps Mars. A clue to the identity of a denser material lies in the high abundance of iron in chondrites, as well as in the existence of the *iron meteorites*, which have densities of 7.5 g/cm^3, approaching that of metallic iron. Nearly as abundant as silicon and magnesium, iron is an excellent candidate for the material that raises the terrestrial planet densities beyond those of silicates. Mercury has by far the largest abundance of iron (most of its mass); the Moon has the least (close to zero). This is interesting in view of the fact that these two heavily cratered planets have similar sizes, the smallest of the terrestrial planets; clearly, their histories and probably their origins were quite different, given the distinct compositions. Nickel also is present in meteorites at roughly 6% of the abundance of iron, and in planets it is expected to be present in similar amounts.

About one-third the mass of Earth is in the form of iron. Much of this may be chemically combined with sulfur or oxygen, and is known to be mostly segregated in a core, for reasons we discuss in section 11.3. Hence Earth is stratified with the densest material toward the center; this is likely to be the case for all of the planets and most of the moons. The outermost chemical layer, or crust, of Earth is composed of minerals containing largely silicon and magnesium with an admixture of lower-density minerals. Aluminum, for example, underabundant relative to silicon and

magnesium but with similar mineral-forming properties, is more abundant in the crust of Earth than throughout the rest of its interior. Venus' strong resemblance to Earth in density and size leads us to conclude that, in bulk composition, it is similar to Earth. Limited chemical measurements of the surface from landed, Russian space probes suggest this to be the case, even though the geologic processes shaping the face of Venus appear to be different from those on Earth (see Chapter 15).

Most of the major moons of the outer solar system, and Pluto, have densities around 2 g/cm^3. This is too low to be accounted for by common silicate minerals, but too high for pure ices. A sensible inference is that these bodies are roughly one-half ice and one-half silicate by mass. Because oxygen is significantly more abundant than carbon, nitrogen, or other ice-forming elements, it is a logical assumption that water dominates the ice component of these moons, with admixtures of ammonia, methane, carbon dioxide, and other ices. Spectroscopic identification of water ice, and of other ices in the cases of Triton and Pluto, seem to confirm these general ideas. Two major moons of Jupiter – Io and Europa – exceed 3 g/cm^3 in density. Io is almost entirely silicate; Europa is lower in density and appears to have a water-ice veneer. They both may have lost water early on, or never acquired significant quantities in the first place.

Interpreting planetary compositions in terms of the internal arrangement, or *structure*, of the various major components is a challenge that requires additional observational tools. We describe the monitoring of earthquakes to infer Earth's internal structure in section 11.3.

11.1.2 The giant planets

Determining the detailed composition of the giant planets, pictured in Figure 11.2, is difficult because of their distance from Earth, and the inaccessibility of their vast interiors. Density can be measured from size and mass, and the values are 1.33 for Jupiter, 0.69 for Saturn, 1.27 for Uranus, and 1.64 for Neptune, all in units of grams per cubic centimeter. These are much lower than the densities of the terrestrial planets plotted in Figure 11.1. Equally important to understanding composition is the determination of the shape of the giant planet and hence its gravitational field. Such information provides constraints on whether especially dense layers are located near the planet's center, and to what extent the outer gaseous layers are pure hydrogen and helium. Measuring the gravitational field requires precise tracking of the orbits of a planet's moons, particularly its closest ones, and this must be done using flyby robotic spacecraft such as the Cassini Saturn Orbiter. Tracking the paths of the spacecraft themselves also yields gravity data. The spin rate of the planet also must be measured, because a faster spin tends to flatten gaseous planets significantly, affecting the distribution of mass in their interiors and hence their gravitational fields.

Both Jupiter and Saturn have such low densities that they must be made up mostly of the light, primordial elements hydrogen and helium; spacecraft spectroscopic measurements show that helium is 10 to 15% the abundance of hydrogen in the outer layers of Jupiter, Uranus, and Neptune, but much lower in Saturn. The giant planets much more closely resemble the Sun in composition than do the terrestrial planets. However, there must be much more of the elements heavier than hydrogen and helium in the giant planets than in the Sun, based on their densities and observed gravitational fields. The composition of the deep interiors cannot be sampled directly but is likely to be largely the abundant rock- and ice-forming elements such as oxygen, carbon, nitrogen, silicon, magnesium, and iron. The tremendous pressures, 70 million times Earth's sea level pressure at the center of Jupiter, force these materials to exist in chemical configurations different from those we are used to seeing on Earth. A schematic slice of Jupiter's interior is shown in Figure 11.3.

Spacecraft and Earth-based identification of hydrogen and helium by spectroscopy and other techniques confirm their presence, at least in the outermost layers. However, other molecules are found to be present, such as methane and ammonia, which likely contain most of the carbon and nitrogen atoms in the outer layers; water is probably present but temperatures in the atmospheres are low enough that it is condensed out below the measurable outer layer of these planets. If Jupiter and Saturn had an overall composition equal to that of the Sun, we would expect that no more than 1% of the mass of each planet would be in the form of heavy elements. However, 10% or more by mass of each planet must be elements heavier than hydrogen or helium, based again on gravity tracking, and this important fact drives scientists to the model described in Chapter 10 in which solid cores form first and then gravitationally attract nebular gas to form the giant planets. In effect, the genesis of Jupiter and Saturn begins with the formation of terrestrial- or ice-moon-type bodies, a process that does not stop until these protoplanets are drowned in hundreds of Earth masses (in the case of Jupiter) of hydrogen-helium gas. Continued infall of icy planetesimals during and after this process "salts" the gas envelopes of these planets with more ice- and rock-forming material.

Uranus and Neptune hold far less hydrogen-helium gas than do Jupiter and Saturn but vastly more than the terrestrial planets. Though not as rich in hydrogen and helium as are Jupiter and Saturn they may contain large amounts of water. Because of their great distances from the Sun, their atmospheres are extremely cold; the water is frozen out of the upper, observable, atmospheres, where instead clouds of methane are seen to form.

Careful measurement of the absorption of sunlight and emission of heat from each of the giant planets reveals that, with the exception of Uranus, each releases more energy in the form of heat than it receives as sunlight. There must be an internal source of energy in each planet, but it cannot be hydrogen fusion, because temperatures and pressures computed for the center of each body are too small to overcome the repulsive forces that prevent protons from fusing together. Even deuterium fusion, which is easier to initiate, cannot be achieved; computations show that a body must be 13 times the mass of Jupiter for such reactions to take place.

The most plausible source of heat comes from the formation of the giant planets themselves. In compressing gas into a self-gravitating, bound sphere, from a state in which the gas originally is spread over a large region of space, potential energy is lost and converted into random energy of motion of the atoms and molecules, that is, into heat. It is the same process that heats the air that you pump into your bicycle tire, but the source of compression is gravitational energy rather than the stored chemical energy in your muscles that you use to move the pump

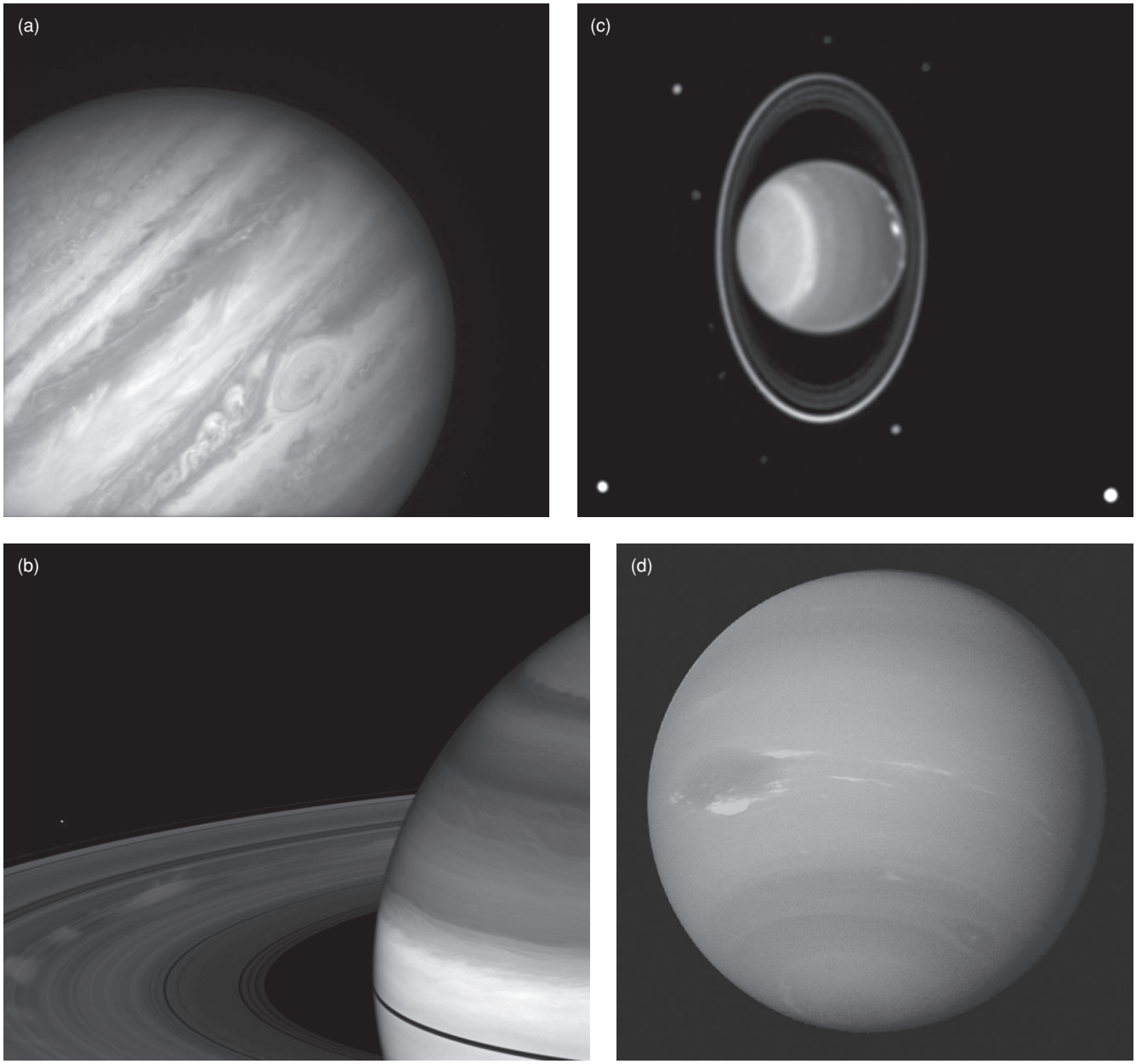


Figure 11.2 The giant planets of our solar system: (a) Jupiter from Hubble Space Telescope; (b) Saturn from Hubble, with contrast exaggerated to show atmospheric patterns; (c) Uranus from Voyager 2, also with enhanced contrast to show very faint banding; (d) Neptune from Voyager 2. Photos (a) and (b) courtesy of NASA and the Space Telescope Science Institute; (c) and (d) courtesy NASA and the Jet Propulsion Laboratory. See color versions of (a), (c), and (d) in plates section.

piston. The initial energy of formation cannot be lost all at once; processes such as conduction, radiation, and convection, which transport heat from the inside of a large body such as a planet, do so in a finite amount of time. Hence heat is still being lost today. University of Arizona scientist W. B. Hubbard and colleagues showed, back in the 1970s, that the excess heat emitted by Jupiter today is consistent with residual heat of formation if Jupiter formed some 4 billion to 5 billion years ago – consistent with the age of the solar system. Neptune's heat yields a similar result.

Saturn and Uranus, however, are a mystery. Saturn emits almost twice as much heat as it should if the energy is that of its initial collapse some 4.5 billion years ago. Is Saturn younger than the other giant planets? A more elegant and sensible explanation comes from the helium abundance, which is depleted in the upper layers as measured from spacecraft. David Stevenson of the California Institute of Technology and Edwin Salpeter of Cornell showed, over two decades ago, that Saturn's interior is cool enough (in contrast to Jupiter's) that helium is chemically insoluble in hydrogen. The helium has been separating out as droplets that fall toward Saturn's center because their density is larger than that of the surrounding hydrogen. This slow rainout of helium contributes additional gravitational energy, which is detected as excess heat emission.

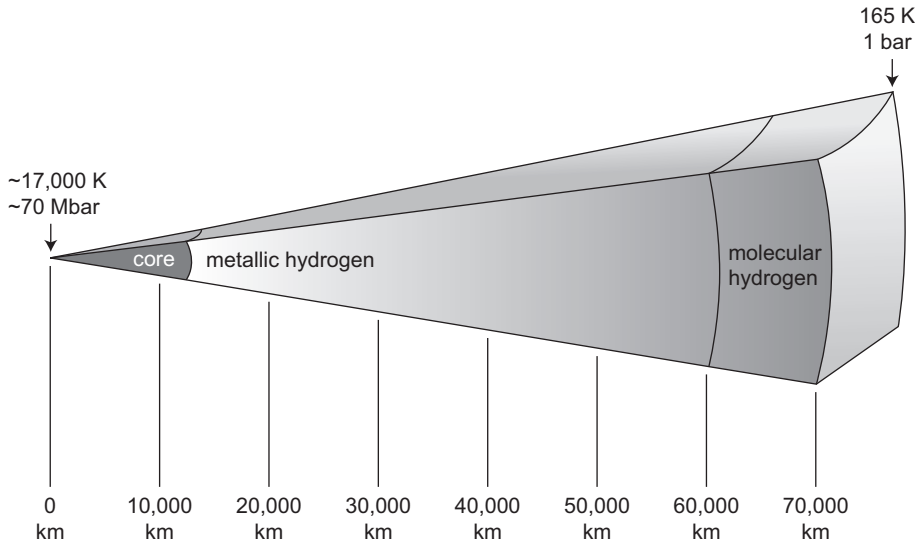


Figure 11.3 A slice through the interior of Jupiter, with distances to the center marked, as well as the pressure and temperature at the top and in the center.

The process of *planetary differentiation*, in which the interior materials are sorted out according to density, is important in Saturn because the ringed planet is less massive than Jupiter, and hence its interior cooled quickly to the point where separation could begin. Subsequent calculations by Stevenson suggested that helium rainout also is occurring in Jupiter, but began more recently than in Saturn. *Voyager* and *Galileo* data on the helium abundances of these giant planets, showing a strong helium depletion in Saturn's atmosphere, but not Jupiter's, appear to support the model.

Measurements showing that Uranus emits essentially no heat, other than what it derives from sunlight, do not yet yield a tidy explanation. Nearly Neptune's twin in terms of size, mass, and density, we expect it to have a similar source of internal heat. However, Uranus spins on an axis that lies parallel to the plane of its orbit around the Sun, rather than close to perpendicular as with Earth and most of the other planets. Over the past decade, Uranus has had one pole tipped toward the Sun, and hence is receiving solar energy in a very different distribution of latitudes than is Neptune. It is possible that this has bottled up or redirected the internal heat of Uranus so that it is not observable. As Uranus moves in its orbit so that the equator, rather than the poles, points toward the Sun, interior energy may be released. That this is happening is suggested by Hubble Space Telescope images of Uranus in the late 1990s showing clouds becoming more abundant on its surface, a surface that *Voyager 2* found to be bland in 1989 (Figure 11.4). Perhaps the "cork" has been popped from the planetary bottle as Uranus moves from a solstice orientation to one of equinox, that is, from summer/winter to spring/fall.

11.2 Internal structure of Earth

We cannot drill more than a few kilometers into the solid Earth with current technologies, yet the center of Earth lies over 6,000 km from the surface. *Apollo* astronauts have drilled only a meter into the Moon, and the *Viking* spacecraft has dug just a few centimeters into Mars. How, then, can we possibly know anything of the internal structure of these planets if drilling significant distances is not possible? Mapping the gravitational fields of the terrestrial planets reveals important information akin to that for the giant planets. Such mapping has revealed the root structure of continents on Earth, information on the patterns of convection (bulk motions removing heat) in Earth's mantle, the nature of the structure beneath the highlands on Venus, and the presence of very dense rock under portions of the Moon's surface.

Earthquakes provide the key to inferring the structure of Earth, the only planet upon which a dense network of seismometers, or earthquake detectors, exists. (A much smaller network, but lacking global coverage, was placed on the Moon during the *Apollo* landings.) An earthquake is an oscillatory movement of the ground, generated by the sudden slippage of rock along a fault (fracture) in Earth's crust. Earthquake waves travel literally around and through Earth, because solid rock is an excellent medium for conducting the wave motion. Rock motions include compression-rarefaction of the rock, or *P-waves*, and shearing motions, or *S-waves*. *P*-waves move easily through solid or liquid media; *S*-waves cannot move through liquid but instead are damped out, because shear forces cannot be maintained in a fluid.

Networks of seismometers recording ground motions were placed around Earth early in the twentieth century, allowing measurement, at many places around the globe of the time for *P*- and *S*-waves to reach various stations from a given earthquake. These times yield the velocity of the seismic waves through the interior and, because *P*- and *S*-waves travel at different speeds, it is possible to precisely locate the geographic point at which an earthquake occurred by measuring arrival times at different stations.

Figure 11.5 illustrates the most important result of such measurements, the inference of the presence of a chemically distinct

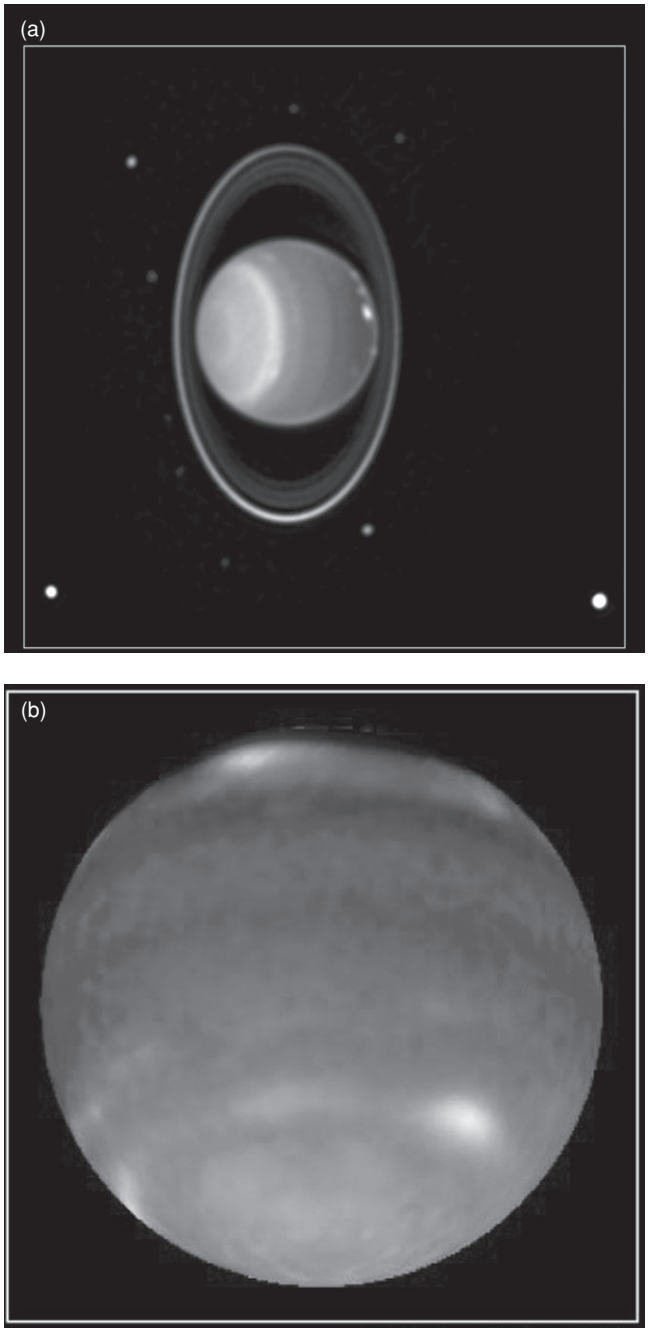


Figure 11.4 Uranus and Neptune from the Hubble Space Telescope, showing continued weather on Neptune (a) and a surge of storms on Uranus (b). Uranus image by Kenneth Seidelman (U.S. Naval Observatory); Neptune images by David Crisp at NASA's Jet Propulsion Laboratory and Heidi Hammel at Massachusetts Institute of Technology Courtesy of NASA and Space Telescope Science Institute. For color version see plates section.

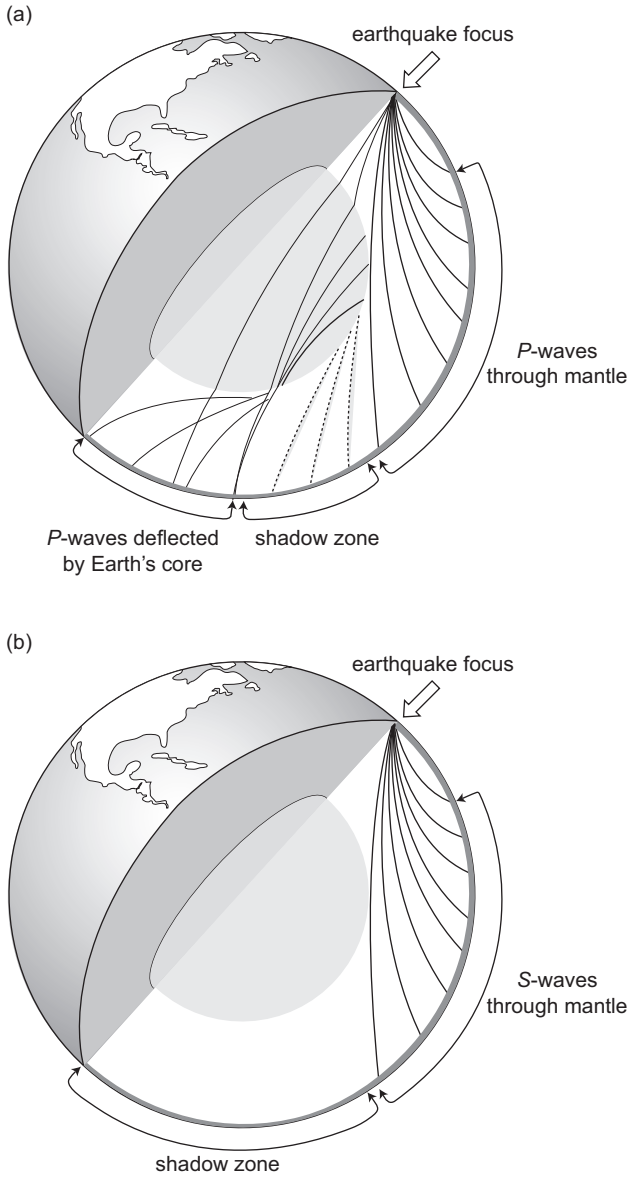


Figure 11.5 (a) P-waves moving through an earth with a central structure different from the bulk outer shell are bent in such a way as to create a shadow zone. (b) S-waves do not propagate at all through a portion of the deep interior, though near the very center, some S-waves do form and propagate a finite distance. Based on Press and Siever (1978) and other sources.

core to Earth. P-waves are observed to be absent from seismometers for certain distances around Earth from the *epicenter* or *focus* of an earthquake, and more highly concentrated elsewhere. This is a strong indication of a structure inside Earth that is bending the paths of, or refracting, the P-waves. The structure could be liquid or solid, but it must have a sharply different density and/or composition from the material above it. S-waves do not accompany those P-waves whose path takes them through this inner structure of Earth, indicating that they cannot propagate through the structure. Hence, there must be a sphere of liquid in the deep interior of Earth – the *core*.

However, the core turns out to be not entirely liquid. The velocities and paths of P-waves are altered in such a way that there must be another structure even deeper in Earth's interior – an *inner core* that is solid and surrounded entirely by the liquid *outer core*.

A promising candidate for a material making up the core is iron, given that the overall density of Earth is significantly higher

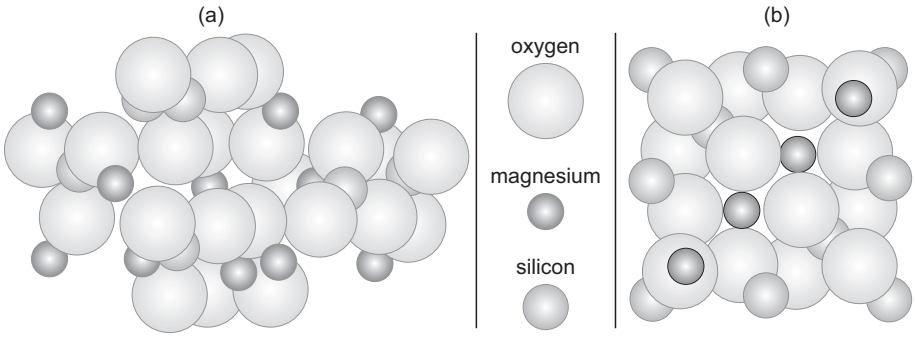


Figure 11.6 Crystal structures of (a) olivine and (b) spinel. After Press and Siever (1978).

than the silicates that comprise our planet's outer layers. Iron is a cosmically abundant element, and meteorites with both a stony and an iron phase suggest that, in larger bodies, the iron separated from the less dense silicates and formed a discrete core under the influence of the gravitational field of the body – a gravitational separation akin to that of helium from hydrogen in Saturn. By estimating the size of the core from the earthquake data and assuming that it is mostly iron, Earth's overall density is reproduced adequately.

However, the core cannot be made of pure iron. Chemically compatible and abundant nickel is likely to be present as well, in small quantities (less than 10%). But more important, the behavior of the core in having a liquid outer region and a solid inner region cannot be reproduced by iron for any plausible temperatures and pressures that are computed for the core. Instead, other elements must be chemically combined with iron in such a way as to allow iron to melt at a lower temperature than it otherwise would, and to provide a means for iron to segregate into a solid inner and a liquid outer core. The velocities of seismic waves in the outer core also argue for the presence of elements in addition to iron.

It is a common chemical fact that combining certain elements or molecules that are compatible in terms of atomic size or valence can lead to the production of liquids at temperatures lower than those for which the pure substances would melt. These are called *eutectic* or *peritectic* solutions. Iron combined with oxygen or sulfur forms such solutions and yields the right melting properties under the tremendous interior pressures of Earth (the central pressure is 4 million atmospheres – 4 million times sea-level pressure on Earth) to account for the dual liquid and solid cores. These chemical, and density, considerations indicate that roughly 10 to 15% of the core is nonmetallic elements: oxygen, sulfur, and perhaps even silicon.

There is much structure near the outer edge of Earth as well. A clear increase in the velocity of seismic waves was established by seismologists at the turn of the twentieth century. The discontinuity occurs some 6 km beneath the ocean floor, but 30 to 40 km below the surface of the continents. This change, known as the *Mohorovičič* or Moho discontinuity (after a Balkan seismologist), defines the boundary between Earth's *crust* and its *mantle*. The boundary is both chemical, in terms of a change in composition, and physical, in that the density changes discontinuously. The chemical change is from crustal rocks that are rich in silicon and aluminum to mantle rocks rich in silicon and magnesium.

Other sharp changes in seismic velocity are seen in the upper 700 km of Earth's mantle. Understanding their origin rests on laboratory simulation of the relevant pressures and temperatures, as well as some theoretical analyses of the chemistry involved. Most notable of these is a sudden drop in velocity at about 70 km indicating that the mantle has transitioned from a stiff upper layer (the lithosphere, which includes the crust of Earth), to a soft, even plastic, behavior (the asthenosphere). Still farther down (400 km) is a sharp increase in velocity where the mineral olivine (a magnesium silicate bearing some of the iron that is not sequestered in the core: Mg_2SiO_4 and Fe_2SiO_4) is forced by pressure to assume a more compact mineral form, called spinel (Figure 11.6). (The iron in the mantle is a residue left behind by the original core formation process; estimates from volcanic lavas generated in the mantle yield only about 10% of the Earth's iron as residing in the mantle and crust.)

At 700 km, pressure forces another phase transition as some of the silicon, magnesium, iron, and aluminum are forced into simpler *oxide* forms: SiO_2, MgO, FeO (*wüstite*), and Al_2O_3. The bulk of the magnesium and iron assume mineral structures called *magnesium silicate* and *iron silicate* (*perovskite*), with the chemical formulas $MgSiO_3$ and $FeSiO_3$. Although the chemical formulas are similar to those of some minerals found in meteorites, the configurations of the atoms are very compact. Below this transition, the mantle is remarkably simple: there is no variation in the depth of layers as there is closer to the surface (that is, no "interior" topography, by analogy with variation of height on Earth's surface), and no phase changes until within a few hundred kilometers of the boundary of the outer core, which resides some 3,000 km below the surface.

Within this transition zone between core and mantle, some remarkable chemistry might be taking place. It has only been possible in the past decade or so to determine the structure of the boundary, using sensitive seismometers on the surface of Earth that determine not only arrival times of *P*- and *S*-waves from earthquakes, but the shapes of the waves as well. Work by geophysicist Thorne Lay of the University of California and colleagues has revealed a complex topography lying on top of the core–mantle boundary. The layer in which this topography resides (called the D″ layer for historical reasons) ranges in thickness from less than 10 km (thinner than this cannot be detected) to several hundreds of kilometers.

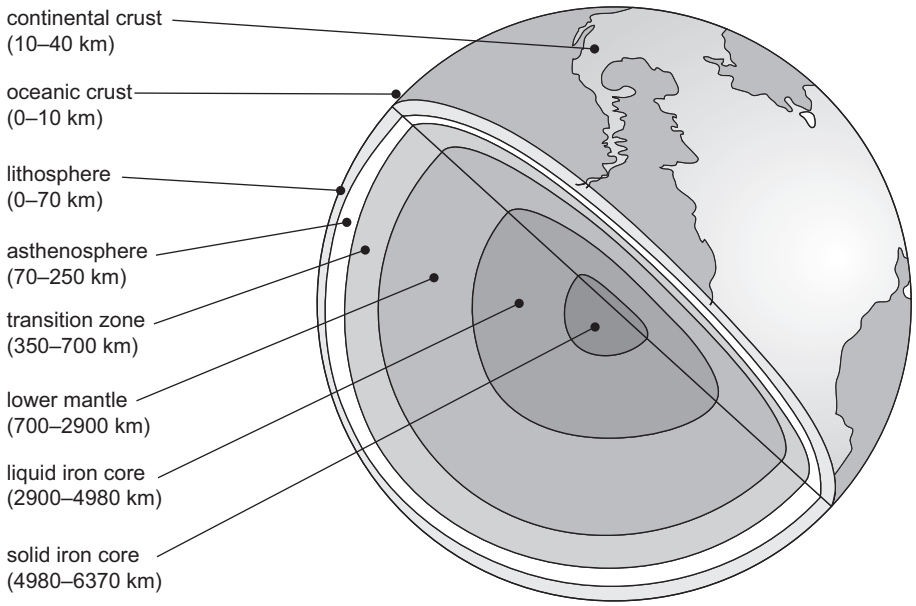


Figure 11.7 Cutaway sketch of Earth's interior from seismic data, laboratory experiments, and theoretical models. The crust is too thin to show at its true scale, and the thickness increases by a factor of four from ocean to continent. The lower mantle is defined as starting around 700 km depth where magnesium and iron perovskites become stable. After Press and Siever (1978).

What is happening within this layer? Laboratory experiments attempt to reproduce the conditions within the D″ layer. Interestingly, liquid iron (representing the outer core) placed in contact with magnesium perovskite (the primary constituent of the lower mantle) reacts chemically to produce the oxides of magnesium, iron, and silicon mentioned above, along with lesser amounts of iron silicide (FeSi). Iron can combine with silicon in this way only because the paired oxygen atom – in wüstite – acquires the electron valence of sulfur, which is just below oxygen in the periodic table of Figure 2.6. The intense pressure at the core-mantle boundary – over one million atmospheres – is enough to distort the electronic structure of the elements.

The technology to simulate these pressures rests on very precisely machined diamonds, whose faceted ends are pushed together by small mechanical presses, with the sample mounted between these ends. A laser heats the sample, and the color of the glowing sample material determines its temperature (see Chapter 3). Such *diamond anvil cells* are in use at laboratories around the world.

Why does the D″ layer exist? One picture is that the liquid iron (with nickel, sulfur, oxygen) of the outer core seeps into cracks in the rock of the lower mantle. Capillary action draws the liquid upward perhaps hundreds of meters into the rock, where it reacts. The oxides and other products of the reaction are much closer in density to the mantle material than they are to the denser liquid metal of the outer core. Hence, any circulation patterns in the mantle will sweep the products of the reaction upward. *Solid-state convection*, in which warm rocks move upward and cooler rocks sink, provides the source of the circulation patterns. Computer models suggest that the upwelling can pile the reaction products into "mountains" hundreds of kilometers high; reactions in sinking regions correspond to "valleys" in the D″ layer.

The overall model for Earth's interior is shown in Figure 11.7. What is striking about this picture of the interior is that most of the chemical and structural complexity of Earth is confined to the crust, where the buoyant distillates of the mantle reside, and the D″ layer. In both cases the interfaces remain poorly understood, and new ideas about the nature of these boundaries will continue to be offered based on seismic measurements and laboratory studies.

11.3 Accretion: the building up of planets

As material is added to the forming planets by collisions (little chunks of rock agglomerating to make bigger rocks), the kinetic energy of impact is converted to heat. An alternative way to look at this is that the potential energy of the dust and small rocks, distributed over a large region of the primordial nebula, is larger than the potential energy of the material gravitationally bound to the growing planet. The lost potential energy shows up as heat, just as it does for the nebular gas going into the giant planets.

The heat added per unit of material (conveniently measured in mass, for example, per kilogram) in the outer layers of the growing planet is equal to the gravitational potential energy lost in each kilogram. The potential energy, in turn, is proportional to the density and radius of the growing planet, which defines the gravitational well into which the material falls. Hence, among the terrestrial planets, Earth and Venus were heated the most, Mercury and the Moon, the least. Two important complications are the average size of the impacting planetesimals at the end of accretion (bigger bodies deposit their energy deeper into the growing planet) and assembly time (longer times allow more heat from the impacts to leak out at the planetary surface, creating lower temperatures overall). These are suggested schematically in Figure 11.8.

Table 11.1 Ionic radii for selected elements

Element	Size[a] (Angstroms)
Be	0.34
P	0.35 (lithophilic)
Si	0.39 (lithophilic)
Al	0.57 (lithophilic)
Li	0.78 (lithophilic)
Mg	0.78 (lithophilic)
Fe	0.82 (siderophilic)
Na	0.98 (lithophilic)
Ca	1.06 (lithophilic)
Sr	1.27 (lithophilic)
O	1.32
F	1.33 (lithophilic)
K	1.33 (lithophilic)
Rb	1.49 (lithophilic)
S	1.74 (chalcophilic)
Cl	1.81 (lithophilic)
Br	1.96 (lithophilic)
I	2.20 (lithophilic)

[a] Ionic radii are given for the element's usual form in Earth's crust.
Data from Broecker (1985) and Mason (1991).

11.4 Early differentiation after accretion

Earth, Venus, and perhaps Mars achieved temperatures throughout parts of their interiors, by virtue of accretion, above the melting point of most silicate minerals. Hence, the earliest Earth had a massive molten region, extending from the surface down partway through the interior. Heat flowed both toward the colder center region, and outward toward space. Elements that previously were bound in the solid crystalline structures of the planetesimals were free in the liquid to rearrange themselves, associating with other compatible elements.

On the basis of their valence structure and the effective size of the atoms, elements can be divided into lithophiles, siderophiles, and chalcophiles. A *lithophile* (from the Greek, "rock-loving") element tends to associate with silicate phases, a *siderophile* ("iron-loving") element with the metal phases, and a *chalcophile* ("ore-loving") in the sulfur-bearing, or sulfide, phases. Chalcophile elements also can be distinguished by their tendency to be volatile and hence to escape from the solid phase. Table 11.1 lists the size of the ions of various elements, where the particular ions chosen are those common in Earth's mantle or crust. Ions significantly larger than the host magnesium or silicon ions have difficulty fitting into the solid crystal structure, and hence tend to stay in the molten rock.

In the early melting of the outer layers of Earth, large ions such as sodium (Na) and potassium (K) tended to reside in the liquid and float to the top of this massively deep *magma ocean*. How much differentiation occurred during the time after Earth reached its present size is controversial, because the precise temperature increase and hence extent of the magma ocean due to accretion cannot be pinned down. Furthermore, Earth's crust has been geochemically cycled and processed extensively in the 4.5 billion years after formation, erasing evidence for an early episode of differentiation. We expect the degree of early geochemical evolution on Venus to be the same as that of Earth, and less on Mars, Mercury, and the Moon commensurate with their smaller sizes. Only Mercury and the Moon are small enough

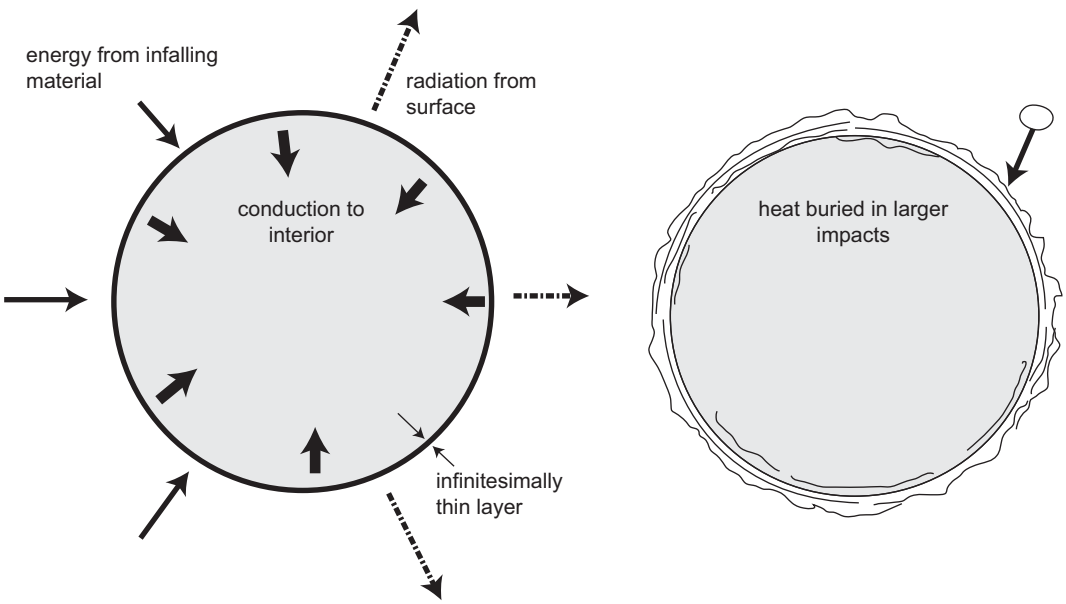


Figure 11.8 Two extreme ways that solid planets can accrete, by small or large planetesimals. On the left, a planet grows by accumulating small grains or boulders, which, as they hit, deposit their heat on the surface. Some of the heat is radiated away by photons; the temperature increase depends on the rate of impacts compared to the rate at which the heat is radiated away. On the right, a planet grows by giant impacts, which gouge out the surface and bury the heat of impact in the planetary interior. The amount of heat that each impact provides depends in this case on the complex details of the impact process. Actual accretion involves both large and small impactors. Adapted from Melosh *et al*. (1993) by permission of University of Arizona Press.

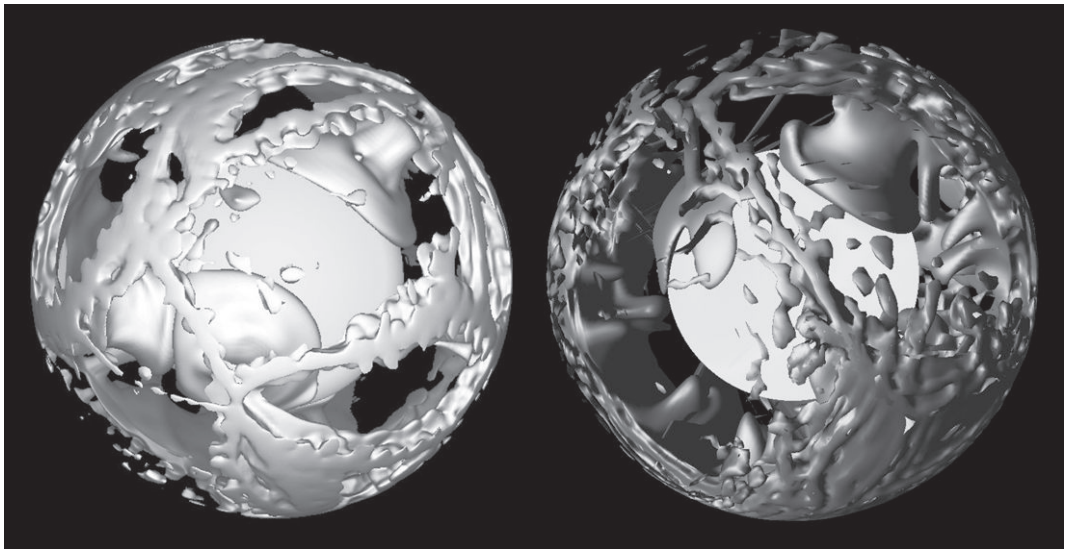


Figure 11.9 Computer model of convection in Earth (Tackley, 1995; Tackley *et al.*, 1994). The model is three-dimensional and includes the presence of the phase transition at the upper–lower mantle interface. The left panel shows hot upwelling currents; the right panel shows cold downwelling currents. The inner sphere, which can be partly seen through the mantle currents, indicates the boundary with the iron core, which convects separately. Figures courtesy of Paul Tackley, University of California at Los Angeles. See color version in plates section.

to have undergone little crustal evolution after the initial accretional episode, and may preserve the chemical evidence of that earliest part of their history. In the case of the Moon, as we discuss in section 11.8, the material out of which it formed may have been derived in large part from an already partly processed Earth. An eventual sampling of the crust of Mercury may be the best way to learn how planetary interiors were altered during the last stages of accretion.

11.5 Radioactive heating

The building blocks of the terrestrial planets were broadly similar, but not identical, to the chondritic meteorites, with more or less of the volatile elements included depending on the distance from the Sun at which particular grains condensed out. Among the elements present were uranium, thorium, and potassium, each of which has isotopes (^{235}U, ^{238}U, ^{232}Th, and ^{40}K) that are radioactive. The half-lives of these isotopes are given in Table 5.1. Interestingly, both uranium and thorium have large ionic radii like potassium, and hence over time have become concentrated in the rocks of the crust, particularly in granite, where the radioactive isotopes of these species average 20 parts per million (ppm) in abundance. This is enough to produce, each year, 0.03 joules of energy in every kilogram of rock. Although this does not seem like much energy (a billion kilograms of granite is required to put out a watt of power), it is still substantial when the entire mass of granitic crust is considered. Roughly 2×10^{22} kilograms of granite are in the crust, leading to a total annual production of energy of 6×10^{20} joules: 20 trillion watts of power.

In the bulk of Earth, the present radioactivity abundances are much smaller; estimates from mantle-derived volcanic rock suggest about 0.1 ppm and an energy production rate at present of 0.0001 joules per kilogram every year. However, the entire mantle, which is roughly 4×10^{24} kilograms in mass (200 times more massive than the granitic part of the crust), generates over 10 trillion watts of power from the decay of radioactive potassium, uranium, and thorium.

At present, then, over half of the heat coming out of continental rock is generated within that rock, with heat from the deeper mantle being the other source. Oceanic crust, however, is depleted by a factor of six from continental in terms of its store of radioactive elements; most of the heat coming from ocean crust had its ultimate origin in solid-state convection in the mantle.

The effect of radioactive heating depends on a planet's size in two ways. The smaller the body, the less radioactive material that is present to heat the interior, and the larger is the ratio of surface area to volume. As a sphere shrinks, the surface area decreases more slowly than the volume. Reduce a planet to half its original size (while retaining the shape of a sphere), and the surface area drops by a factor of four while the volume (and hence the number of radioactive atoms within the planet) drops by eight. Since more relative surface area allows faster cooling, smaller objects cool more quickly than bigger ones. Based on relative sizes, Venus' thermal history was similar to Earth's, but Mars likely cooled more quickly than Earth, and Mercury even more rapidly. We see the evidence for this in the heavily cratered surface of Mercury and in the bimodal nature of Mars, wherein both massive volcanoes and heavily cratered terrains exist.

Sufficient heating is occurring today in Earth's mantle to soften the rock and allow bulk flow to remove the heat. The core of Earth is releasing heat to the mantle as well, so that the nature of the heat flow is somewhat complicated (Figure 11.9). Simple patterns of bulk convective motion of the mantle are interrupted by plumes of hot material driven by heat from the core. These deep-seated plumes may reach the surface in the form of large volcanoes, which are then dragged laterally by plate motion to form island chains such as Hawaii.

11.6 Formation of an iron core

No more than a few tens of millions of years after the Earth began to grow toward its present size, temperatures throughout the deep interior were enough to partially melt the mixed solids of silicate, and iron. Iron melts at a temperature a couple of hundred degrees below the melting point of the major silicate component, magnesium silicate, and would be expected to sink to the Earth's center by virtue of its higher density. However, a plausible mechanism for iron core formation requires that a substantial fraction of the silicates melted as well, to allow the denser iron to separate readily from the surrounding material and sink. Because the iron core formation involves taking denser material from a distributed state and placing it in the very center of the planet, gravitational energy is released. The sinking of helium to the center of Saturn, creating extra heat from gravitational energy, is an entirely analogous process discussed earlier in the chapter. The total iron content of Earth corresponds to 32% of the mass of our planet, and the density of iron is about 50% higher than that of silicates, and so, the differentiation process releases an amount of heat not very much less than the total accretion energy of Earth; this undoubtedly helped to ensure melting of Earth's upper layers at that time.

The iron core is able to generate a magnetic field. As the core convects to remove heat to the cooler mantle layers above it, the motions of the electrically conductive iron have the potential to induce magnetic fields. Schematically, if a "seed" magnetic field is initially present in the core (left over from magnetic fields in the solar nebula that magnetized rocks and iron grains), then the moving fluid generates electric currents, which in turn generate a stronger magnetic field. This self-perpetuating process, energized by the heat slowly leaking from the core, is called a *magnetic dynamo*.

When did core formation occur? Theoretical calculations suggest that temperatures were high enough to initiate mantle melting during accretion, but it is important to have an independent constraint on the time of initiation and duration. Isotopes of lead provide that determination. The element lead is chemically compatible with iron and hence followed iron into the core. Uranium, on the other hand, tended to stay in the crust and mantle. Heavy isotopes of lead (^{206}Pb and ^{207}Pb), however, are daughter products of uranium decay, with long half-lives (4.5 and 0.7 billion years, respectively). Thus, by measuring the abundances of these daughter isotopes we have a potential way of determining when the core separated from the mantle. Ancient lead-bearing rocks on Earth's surface are compared with lead isotopic abundances in meteorites to infer that core formation occurred during the first few tens of millions years of Earth's history, essentially coincident with the late stages of accretion. A check can be made using xenon isotopes, which corroborate this determination.

11.7 Formation of the Moon

The origin of the Moon has always been a difficult issue because our natural satellite is unusually large relative to its primary (Earth) and resides in a circular orbit. Capture of the Moon after its formation is possible but extremely unlikely, requiring just the right set of conditions; capture into a tight circular orbit (the Moon's orbit has been slowly evolving outward with time because of the dissipating effects of ocean tides) is even more improbable. Formation in place at the same time as Earth also has difficulties when one tries to model the process by computer. Finally, fission, wherein a rapidly spinning molten Earth split off the Moon, also has some problems with physical plausibility, but neither this nor formation in place could be ruled out completely on theoretical grounds.

The *Apollo* missions to the Moon returned rock and dust samples that virtually eliminated all three models considered above. In spite of the Moon's small size, and hence limited geologic activity over time, the rocks were more typical of Earth's mantle than of primitive meteorites. However, even more chemical processing beyond that of Earth's mantle was implied: the rocks were strongly depleted in certain elements as volatile or more so than potassium, relative to those of Earth's mantle. In a very crude sense, one could obtain lunar material by taking terrestrial mantle rocks, heating them to temperatures at which they could vaporize, and recondensing only the less volatile constituents. (The term "very crude" must be taken literally, because the described process does not fully explain the lunar composition.)

This geochemical puzzle prompted planetary scientists in the mid-1980s to consider that the Moon might be the product of a huge collision between Earth and another planet-sized body: *a giant impact*. Conditions in the early solar system were right for such an impact. Early on, planetesimals were small and were in roughly circular orbits, which resulted in gentle collisions, and hence sticking or accretion. As planets grew from planetesimals, close passes of bigger bodies altered orbits to make them elliptical, and hence increased relative collision speeds. By the time the terrestrial planets were formed, encounter velocities with solar system debris, on highly elliptical orbits, ensured catastrophic collisions in most cases. This was the case both in the inner and the outer solar system: the newly formed giant planets stirred up nearby planetesimals and ejected them into distant orbits, which we recognize today as the cometary Oort Cloud. The rate of impacts on planets decreased exponentially with time over the first few hundred million years of solar system history, as debris was swept up or ejected (see Chapter 7).

Small bodies hitting big ones would vaporize and melt, disseminating their products in the crust of the big bodies. Big bodies hitting other big bodies could have more devastating consequences. A giant impact with Uranus likely tipped that planet on its side and spun out a disk from which its moons formed. Detailed computer simulations show that a planet one to several times the mass of Mars striking the Earth could have spun off a large amount of the Earth's mantle, very little iron core, and a fraction of that debris would have entered circular orbit around Earth while the remainder was lost into orbit around the Sun or reaccreted onto Earth (Figure 11.10).

Much of the material that shot into orbit was vaporized, with only the least volatile material remaining solid. Some recondensation occurred, but in the absence of a nebular gas providing the conditions for full recondensation, much of the volatile material (water and the volatile lithophilic elements) was lost. Absence of debris from Earth's core resulted in little iron being present, and the Moon's present density is consistent with little or no iron. Accretion of the material in circular orbit to form the Moon was apparently enough to cause melting of the upper 500 km or so

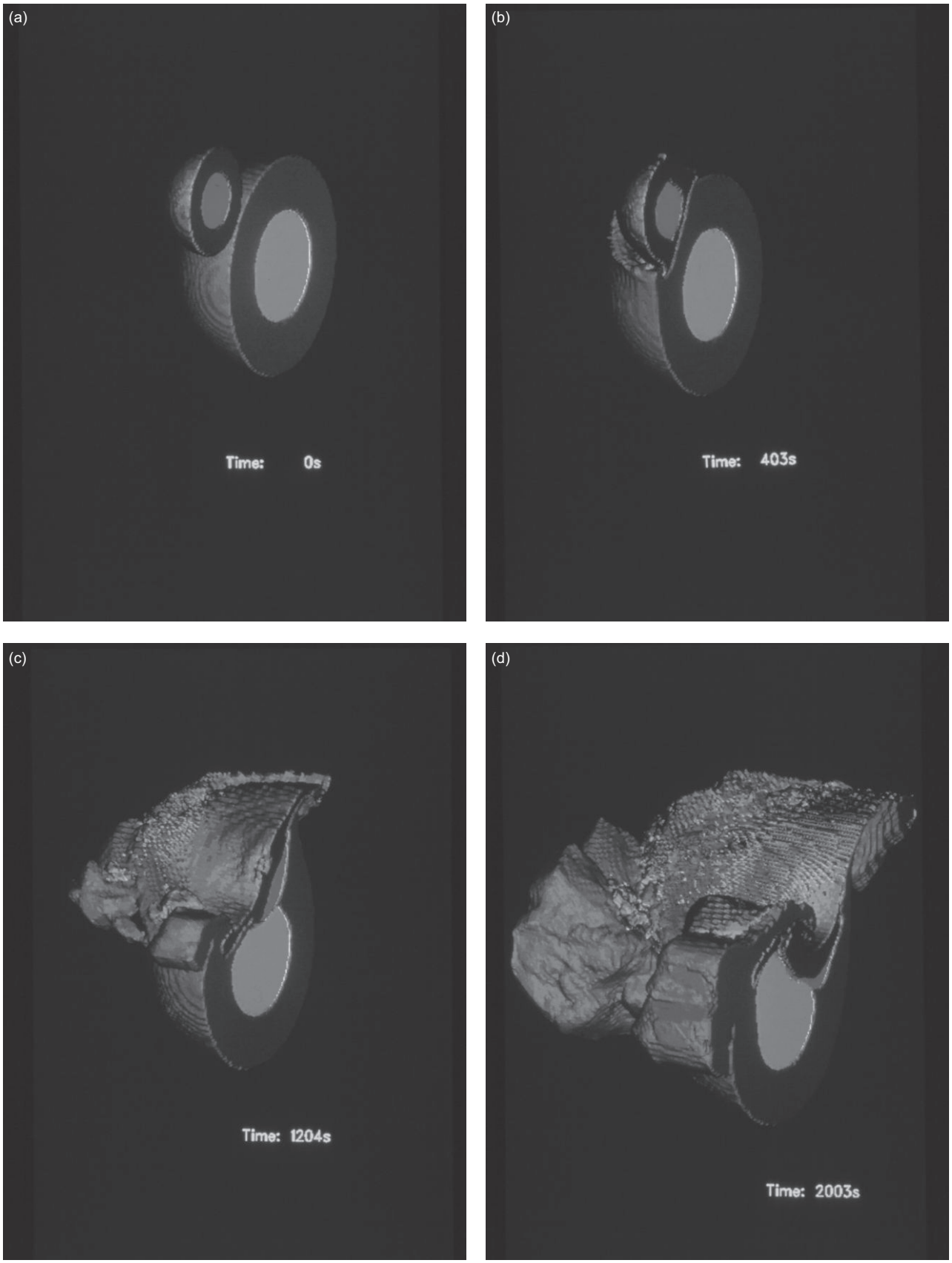


Figure 11.10 Computer calculations by M. E. Kipp (Sandia National Laboratories) and H. J. Melosh (The University of Arizona), showing early stages in the formation of the Moon as a Mars-sized planet strikes Earth. Both Earth and the impacting planet are shown sliced in half so as to reveal what is happening in the interiors. The iron-rich core can be seen as an inner circle in each planet prior to impact. Compared to the mantle of Earth, the core is hardly disrupted. Elapsed time is shown on each panel. Images courtesy of H. J. Melosh. See color versions in plates section.

of Earth's new satellite, because geochemical analysis indicates that the lunar surface is strongly enriched in lower density minerals that likely floated to the top during a molten phase. The ancient lunar highlands are especially enriched in these minerals. Higher density minerals that resemble basalts on Earth have flooded large basins on the Moon, forming the *mare*.

When did the Moon's formation from Earth occur? The oldest lunar rocks found, from the highland provinces, date by radioisotopic techniques (Chapter 5) at 4.4 billion to 4.5 billion years ago; certainly the Moon is no younger than this. This also sets a limit on the time when the Earth's core formed: it had to be before the lunar-forming impact because the Moon is so depleted in iron. Most likely is that the lunar impact occurred extremely early in Earth's history, close to or before 4.5 billion years ago. Earth was not a single planet for very long. Venus, on the other hand, does not possess a moon, and hence either never suffered a giant impact or experienced one that left it in retrograde rotation without a companion, in which case, the ejected material was either reaccreted or lost to solar orbit. Pluto has a moon, Charon, that is even closer in mass to its primary than is the Moon to Earth. It may have formed from a large impact on Pluto, probably by another large Kuiper Belt object whose orbit was stirred up by a close pass to Uranus or Neptune.

What was the origin of the impactor that struck Earth? This remains a mystery, but it is clear from the geochemistry of the Moon that the impactor had to have had a composition similar to that of Earth. Because some of its mass went into the debris that formed the Moon, gross compositional differences would show up in the lunar rocks. Because those rocks do so closely resemble a devolatilized Earth's mantle, the impactor could not have been very different from terrestrial mantle composition.

Figure 11.11 summarizes the timescales for the earliest events in Earth's history, up through core formation. The enormous upheavals in the first 2% of Earth's history, in large measure, are a reflection of the crowded solar system environment at the time: the final stages of growth of Earth by sweep-up of smaller debris heated the planet to high temperatures (with a contribution from internal radiogenic elements as well), and the apparent presence of large bodies in eccentric orbits that crossed those of the planets set the stage for the catastrophic collision that led to lunar formation.

11.8 Origin of Earth's atmosphere, ocean, and organic reservoir

Earth's earliest atmosphere was a cloud of silicate vapor surrounding it during its accretion and core formation. As accretion stopped and core formation ended, the surface cooled and the silicate vapor condensed to form molten and solid rock. If this process concluded early enough, and this is uncertain, Earth would have been surrounded by a remnant primordial atmosphere of molecular hydrogen and trace amounts of other gases. This primordial atmosphere very quickly was swept away by the strong solar wind and is of little consequence to the rest of Earth history.

From whence came the gases that made up the "permanent" atmosphere? Outgassing from Earth's interior, of trace gases trapped in rocks, could have put hydrogen sulfide, carbon dioxide, and a large amount of water (all originally dissolved in the early magma ocean) in the atmosphere and on the surface. The origin of these volatile materials may not have been the vicinity of the forming Earth – where temperatures were too high to condense water – but instead may have been farther out in the forming solar system. Impactors that came from the outer solar system – comets – were rich in water ice, organics, carbon dioxide, carbon monoxide, and ammonia.

The comets are detritus from the formation of outer solar system bodies. Although hundreds of earth masses of comets now reside in orbits far from the Sun, early in the history of the solar system comets were more commonly in orbits that intersected the orbits of Mars, Earth, and Venus (based on computer studies of solar system formation). Collisions of comets with the planets would have released the cometary ices and gases into the atmospheres of the target planets. Early in Earth's history, the first couple of hundred million years, cometary material including water might have been episodically added to the atmosphere. However, the ratio of deuterium to hydrogen (D/H) in the water ice portion of comets is twice that in ocean water on the Earth. No plausible way has been found to lower the value after it has been added to the Earth. Therefore, comets do not appear to be the primary source of Earth's water.

Two alternative possibilities have been proposed. Bodies in the asteroid belt would have been richer in water than material near the Earth, and as discussed in Chapter 10, Jupiter perturbed that material into orbits that could have allowed accretion by the Earth. Most of this material would have been in the form of bodies as large as the Moon or even Mars, so that these collisions would have been violent. Nonetheless, the net affect would have been the addition of water to the growing Earth. Carbonaceous meteorites, some of which may have been derived from the asteroid belt, have a D/H range that averages out to the value present in the Earth's oceans. However, some of the details of the elemental and isotopic abundances in the carbonaceous chondrites limit to 1% the amount of this material that could have been added to the Earth. It is possible that other types of chondrites were present in the asteroid belt that today are poorly known, such as a new class of bodies represented by a handful of so-called "main belt comets", but for the moment this is speculative. Alternatively, water could have been adsorbed on rocky grains closer to the Earth, and brought in through a gentle rain of this material. While laboratory studies show that enough water might have stuck to the grains to explain the abundance of the Earth's oceans, the presence of such a water-laden dust layer in the nebula remains speculative.

Even if comets were not the source of the Earth's water, comets probably brought in carbon dioxide, carbon monoxide, methane, ammonia, nitrogen, and other gases. Carbon dioxide also could have been available from rocks in Earth's mantle, and the early atmosphere likely was dominated by this gas after condensation of water. Molecular oxygen is essentially nonexistent in comets, is nearly absent from Mars and Venus, and was absent from the early Earth atmosphere. That this is so is demonstrated in part by minerals in ancient rocks that would have been unstable in an atmosphere composed of oxygen (Chapter 17).

As described in Chapter 10, Jupiter played the key role in perturbing the orbits of bodies in the asteroid belt allowing for a number of these to collide with the growing Earth. However, all of the giant planets, especially Jupiter, also were very effective

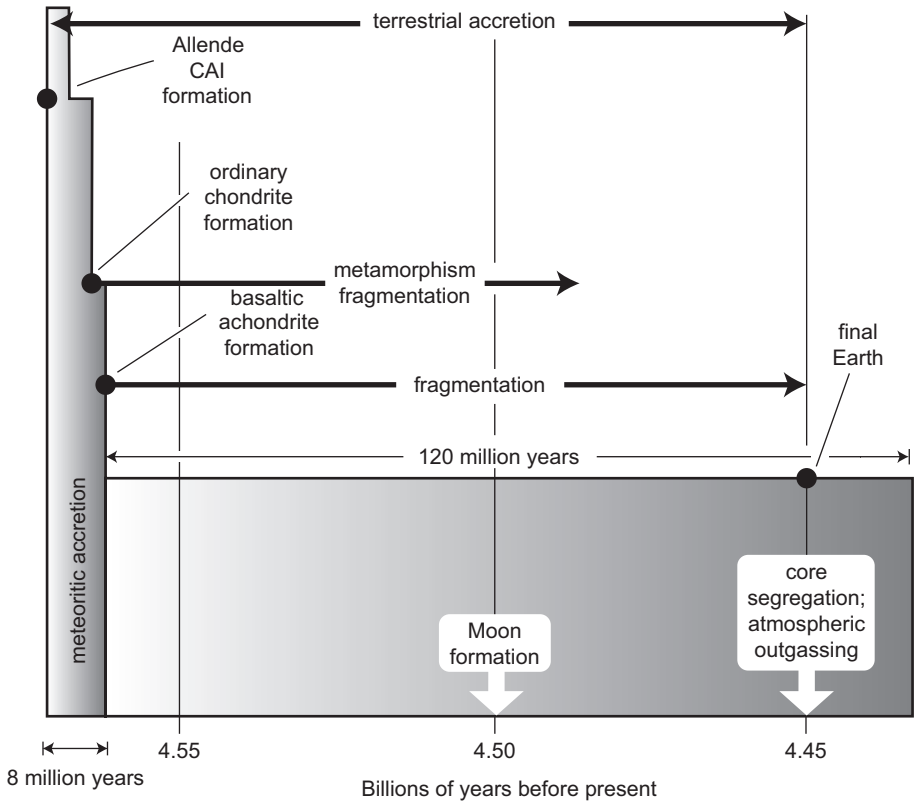


Figure 11.11 Timescales for the formation of Earth and early events in its history, as developed by Claude Allègre and colleagues from radioisotopic analyses of meteorites and lunar rocks. "Allende CAI" refers to particular phases in the Allende meteorite that predate formation of the bulk portion of the chondrites. "Basaltic achondrites" are a class of meteorites that have undergone chemical differentiation and hence are less primitive than the chondrites. Redrawn from Allègre *et al.* (1995) by permission of Elsevier Science Ltd.

in clearing the solar system of planetesimal debris, with much of the material being ejected permanently into distant orbits, or forced into the inner solar system where the icy material collided with the terrestrial planets. Had the giant planets not swept the solar system clear, the impact rate in the inner-planet region might have remained high for billions of years, making for an unstable environment on Earth and frustrating the earliest origin and survival of life.

Although atmosphere-supplying impactors hit Earth with high velocity, much of the material may have fragmented in the protoatmosphere and reached the surface at low speeds. A significant portion of the organic molecules present in the comets and meteorites may have survived intact to the surface. Thus, the early ocean likely was seeded with large amounts of organic compounds, with complexity up to and including amino acids, the building blocks of proteins (see Chapter 12), which have been found in meteorites. As the impact rate declined and Earth's surface began to stabilize, the materials necessary to initiate a biosphere were very likely in place.

11.9 The Late Heavy Bombardment

Evidence primarily from the lunar cratering record indicates that, somewhere between 3.8 and 4.1 billion years ago, a dramatic increase occurred in the rate of impacts. While the evidence for such a period of enhanced bombardment seems solid, its explanation has been elusive. A group of dynamicists from the Observatoire Côte d'Azur in Nice, France, have come up with an explanation that also serves to explain the distribution of orbits of bodies in the Kuiper Belt. In their model, which has come to be known as the "Nice model", the giant planets initially formed much closer to each other than they are today, with Neptune at only 17 AU instead of 29 AU, and Jupiter at 5.5 AU instead of 5.2 AU. This configuration was stable for a few hundreds of millions of years, but interactions between the giant planets and the disk of solid debris they were progressively ejecting from the solar system, along with interactions between the planets themselves, led to small shifts in their orbits. At some point, the orbits of Jupiter and Saturn were such that Saturn's orbit period was just twice that of Jupiter: a so-called 2:1 resonance. This led to much stronger gravitational interactions among them, making the orbits of Jupiter and Saturn eccentric and pushing Uranus and Neptune outward to their current orbits (Figure 11.12). The rate of scattering of solid debris both inward toward the terrestrial planets and outward increased dramatically, and the rate of impact cratering dramatically increased in the region of the terrestrial planets. The timing of this dramatic event is not precisely fixed by the model but plausibly corresponds to that of the Late Heavy Bombardment. Slight differences in initial conditions in the models lead to dramatically different details – in one case, Uranus and Neptune switch places – but the general result of increased scattering of debris toward the terrestrial planets seems a common outcome. While the Nice model is only a model, observations of the configurations of giant planets in other planetary systems

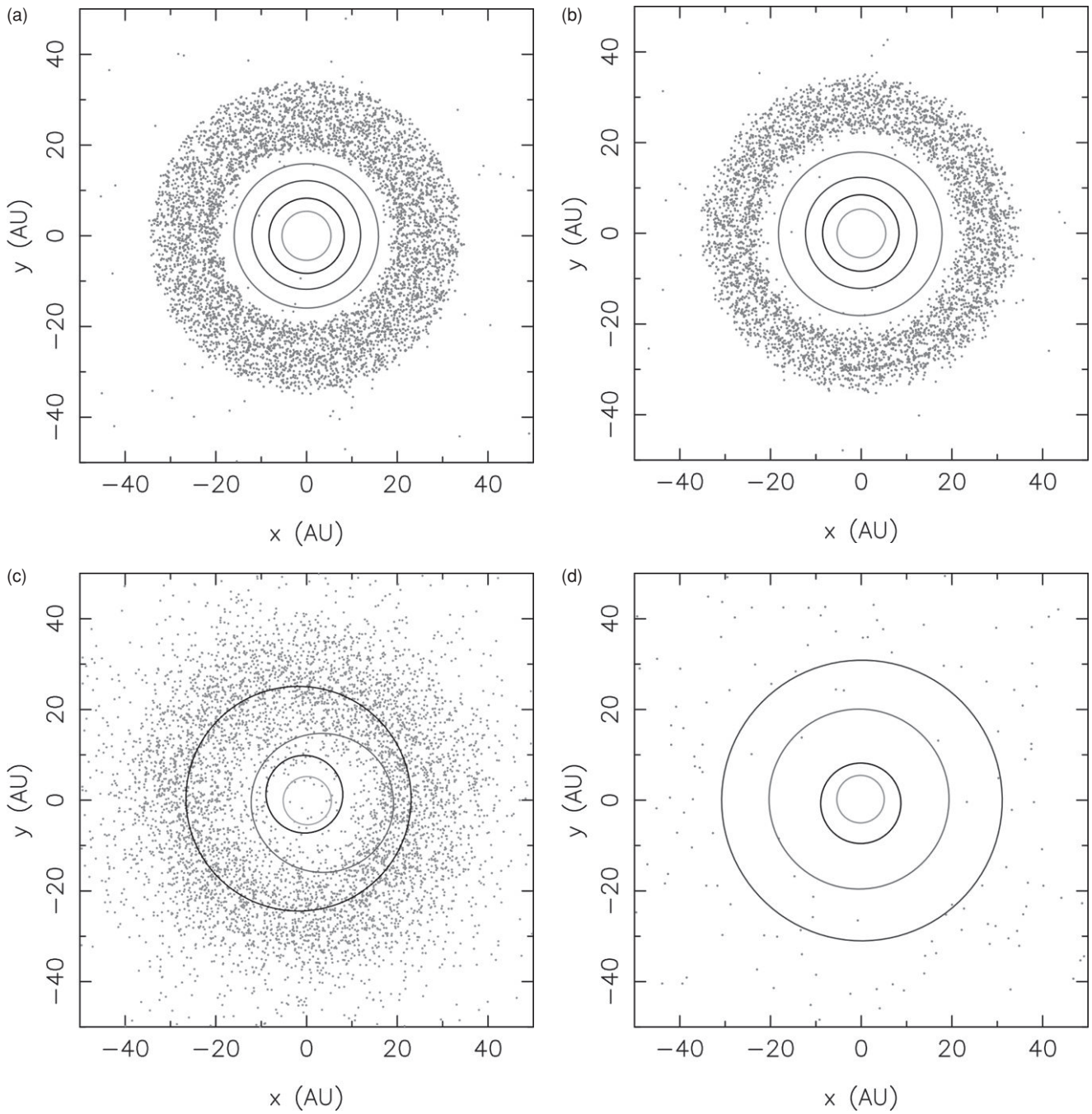


Figure 11.12 Depiction of events associated with the passage of Jupiter and Saturn through the 2:1 resonance of their orbits, according to the Nice model. The orbits of the giant planets are shown as ellipses looking down on the solar system (a) before, (b) at the onset, (c) during, and (d) after the passage through resonance. From Gomez *et al*. (2005).

should give us a perspective on whether such dramatic events might indeed occur early in the history of planetary systems.

11.10 From the Hadean into the Archean: formation of the first stable continental rocks

Even after the early earth crust stabilized, continuing impacts and the vigorous convective activity of Earth's mantle discouraged the preservation of the crustal material over time. The early crust may have had a composition somewhat similar to present-day oceanic crust, depleted in magnesium relative to the composition of the mantle. Continental-type crust required repeated cycling of crustal basalts, with separation of silicon and other elements from the magnesium; such crust was later in coming (see Chapter 16).

The oldest whole rock samples on Earth date back almost 4.0 billion years. These ancient rocks, seen in northern Canada, are composites of *mafic* (magnesium and iron-rich) and *felsic* (less iron- and magnesium-rich, more abundant in silicon) rocks. The former are typical of oceanic basalts, the latter of more continental-type rocks. The samples show evidence of having

been metamorphosed (subjected to episodes of modest pressure and high temperature) in a way that suggests processing in and beneath a primitive basaltic crust. Also present in these rocks are rounded pebbles that appear to be sedimentary, that is, laid down in an environment containing liquid water. Belts of these rocks appear to be the remnants of the earliest continents. They indicate that continental-type crust, floating buoyantly atop a denser mantle, began to appear about 500 million years after the formation of Earth; whether continents could have formed much earlier is unknown. The chemistry of oceanic and continental rock formation is explored in more detail in Chapter 16.

This Hadean Earth, while vastly different from the present planet, set the stage for what was to follow. By 3.8 to 4.0 billion years ago, the growth of continents, the stabilization of liquid water, and the decreasing impact rate made for an increasingly predictable and benign environment. Increasing environmental stability characterized the transition from the Hadean era to the Archean eon of Earth.

Summary

The Hadean era of the Earth spans the time from formation to the presence of the first whole rocks in the geologic record. This is therefore the era in which information on the state of the Earth must be derived from meteorites, from the Moon, and from modeling of planetary processes. The planets of the solar system can be divided according to their density into the solid terrestrial planets, made mostly of rock and metal, and the giant planets, made mostly of hydrogen and helium. Uranus and Neptune are distinguished from Jupiter and Saturn by having far less hydrogen and helium, and proportionately more water. A third class of bodies is made of various proportions of water ice and rock (plus metal); these are the icy moons of the outer solar system, and dwarf planets like Pluto and other Kuiper Belt objects. The Earth's internal structure, revealed through careful measurement of seismic waves propagated by earthquakes, includes a chemically distinct core that is divided into an outer liquid and an inner solid core. The core is mostly iron and other metals with an admixture of oxygen or sulfur. Above the core is the mantle, which itself may be layered chemically, but is made largely of silicates. It is solid, but flows slowly in the same manner as glass does in very old windows. At the core–mantle boundary a complex mixing of the molten iron and solid silicates may be taking place. Above the mantle is the solid crust of the Earth, another chemically distinct layer rich in silicon and aluminum compared to the mantle. The growth of the planets and their moons by addition of material resulted in the release of heat, leading to substantial melting of their interiors. In the case of the Earth, collisions with lunar- and Mars-sized bodies occurred multiple times during its growth, the last of which was a glancing blow that enabled material to remain in orbit, forming the Moon. Meanwhile in the outer solar system, the giant planets may have been spaced more closely together than they are now, orbiting between 5.5 and 17 AU. However, interactions with the remnant disk of debris, and between the giant planets themselves, could have led to a dramatic reshuffling of orbits that is seen in the lunar cratering record as the "Late Heavy Bombardment". The cooling of the Earth after its formation continues to the present, with heat transported from the interior not only from the energy of formation but also from the decay of radioactive elements that progressively became concentrated in the crust.

Questions

1. Some meteorite properties suggest that rocky bodies were strongly heated by ^{26}Al, a very short-lived radioisotope of aluminum. How might the asteroids help determine whether this heating actually occurred? What would you look for?
2. Calculate the temperature rise associated with the formation of the Earth's iron core, assuming that the iron started out fully mixed with the silicates throughout the Earth (this is an oversimplification of what happened, but one still derives a useful number).
3. What might have been different about Earth's Hadean and Archean history had the Moon not been present?
4. Go online to the exoplanet encyclopedia (http://exoplanet.eu/) and examine the orbits of planets in multiple planet systems. Do the configurations you find there seem to argue for or against, or are neutral with respect to, the Nice model?

General reading

Broecker, W. S. 1985. *How to Build a Habitable Planet*. Eldigio Press, Palisades, NY.

Cloud, P. 1988. *Oasis in Space: Earth History from the Beginning*. W. W. Norton, New York.

Press, F. and Siever, R. 2001. *Understanding Earth*. W. H. Freeman, New York.

References

Allégre, C., Poirer, J.-P., Humler, E., and Hofmann, A. W. 1995. The chemical composition of the Earth. *Earth and Planetary Science Letters* **134**, 515–26.

Gomes, R., Levison, H. F., Tsiganis, K., and Morbidelli A. 2005. Origin of the cataclysmic Late Heavy Bombardment period of the terrestrial planets. *Nature* **435**, 466–8.

Jeanloz, R. and Lay, T. 1993. The core-mantle boundary. *Scientific American* **268**(5), 48–55.

Mason, S. F. 1991. *Chemical Evolution*. Clarendon Press, Oxford.

Melosh, H. J., Vickery, A. M., and Tonks, W. B. 1993. Impacts and the early environment and evolution of the terrestrial planets. In *Protostars and Planets III* (E. H. Levy and J. I. Lunine, eds). University of Arizona Press, Tucson, pp. 1339–70.

Owen, T. and Bar-Nun, A. 1995. Comet, impacts and atmospheres. *Icarus* **116**, 215–16.

Press, F. and Siever, R. 1978. *Earth*. W. H. Freeman and Company, San Francisco.

Spudis, P. D. 1992. Moon, geology. In *The Astronomy and Astrophysics Encyclopedia* (S. P. Maran, ed.). Van Nostrand Reinhold, New York, pp. 452–5.

Squyres, S., Reynolds, R. T., Cassen, P. M., and Peale, S. J. 1983. Liquid water and active resurfacing on Europa. *Nature* **301**, 225–6.

Tackley, P. J. 1995. Mantle dynamics: influence of the transition zone. *Reviews of Geophysics* **33** (Suppl.), 275–82.

Tackley, P. J., Stevenson, D. J., Glatzmaier, G. A., and Schubert, G. 1994. Effects of mantle phase transitions in a 3-D spherical model of convection in the Earth's mantle. *Journal of Geophysical Research* **99**, 15,877–901.

Taylor, S. R. and McLennan, S. M. 1995. The geochemical evolution of the continental crust. *Reviews of Geophysics* **33**, 241–65.

Weissman, P. 1992. Comets, Oort cloud. In *The Astronomy and Astrophysics Encyclopedia* (S. P. Maran, ed.). Van Nostrand Reinhold, New York, pp. 120–3.

12

The Archean eon and the origin of life
I Properties of and sites for life

Introduction

The close of the Hadean and opening of the so-called Archean eon is defined and characterized by the oldest whole rock samples found on Earth, 4.0 billion years old. At the opening of the Archean, Earth had an atmosphere rich in carbon dioxide, with perhaps some nitrogen and methane but little molecular oxygen, and liquid water was stable on its surface. Mantle convection had begun producing oceanic basalts and continental-type granitic rocks. The rate of impacts of asteroidal and cometary fragments had decreased significantly. The Moon, formed from Earth at the end of accretion some half billion years before, could be seen in the terrestrial sky.

By 3.5 billion years ago, rocks were present that record definitive evidence for life; more controversial evidence exists back to almost 3.9 billion years. Large sedimentary or layered formations in ancient limestones contain concentric spherical shapes, stacked hemispheres and flat sheets of calcium carbonates (calcite), and trapped silts. These *stromatolites* are best understood as the work of bacteria from 3.5 billion years ago, precipitating calcium carbonate in layers as one of the byproducts of primitive photosynthesis. (Present-day active stromatolite-forming colonies can be found in Shark Bay, Australia.) If the interpretation is correct, life on Earth was present then and somewhat earlier as well, because such bacteria constitute already reasonably well-developed organisms.

It therefore appears that, as Earth settled down from the chaos of accretion, core formation, and impacts, life was able to exist on its surface (Figure 12.1). The same might be true for Mars, but the evidence discussed later in the chapter is vague and controversial. How did life arise on the Earth? Could it have arisen on the neighboring planets as well? Is there life in other planetary systems? Why was Earth able to sustain life over billions of years of change, and the other terrestrial planets not? How did life alter the Earth environment?

These are questions whose explorations constitute the remainder of the book, including Part IV, where human kind's role is examined. In the present chapter, we outline the definition of life and the essential structures that make it possible.

12.1 Definition of life and essential workings

12.1.1 What is life?

No completely satisfactory definition of life – or of "living things" – has yet been devised. Most simple definitions of life – something that grows spontaneously, or something that replicates itself – fail because they either include demonstrably nonliving things or exclude certain particular living organisms. Crystals such as snow or pyrite grow but are not biological in nature; offspring of separate but related species such as mules (offspring of a donkey and a horse) are almost invariably unable to reproduce, yet clearly are living.

Some biologists lean toward a definition that incorporates the concept of *Darwinian evolution*, defined broadly to mean reproduction, variation of characteristics from one generation to another, and natural selection whereby some individuals with specific traits gain an advantage over others and hence are more successful in producing offspring. In this context, one working definition of life might be "a self-sustained chemical system capable of undergoing Darwinian evolution," as devised by University of California biologist Gerald Joyce and colleagues.

There are two major drawbacks to this definition. First, it has become clear that, although species do evolve, the classical Darwinian concept of natural selection is only one factor that comes into play in such evolution. Second, the definition may

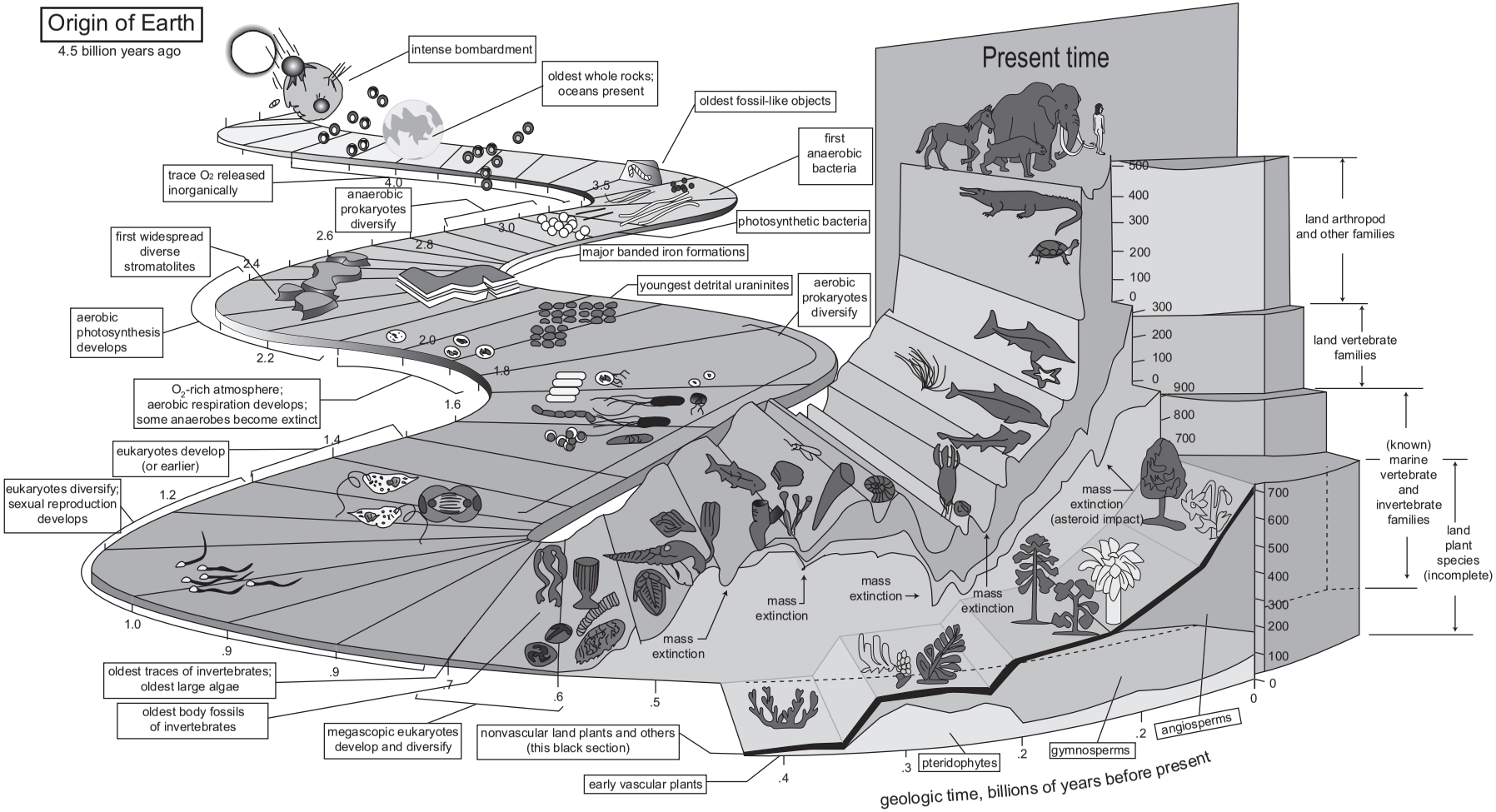


Figure 12.1 Schematic history of life on Earth, showing where key milestones in the history of life likely occurred relative to geologic events on Earth. Beginning at the Vendian–Cambrian diversification of life (Chapter 18), the rise and fall with time of the number of families of land and marine creatures is depicted.

be unnecessarily narrow in that "life" on other planets might not undergo Darwinian evolution, but might still involve biochemical reactions resembling those on Earth; non-Darwinian evolution might have occurred in the very earliest, primitive organisms on our planet as well. The definition also excludes "artificial life," experiments in computer information replication described in Chapter 13, but could easily allow inclusion of such experiments by replacing the phrase "chemical system" by "material system," as has been suggested by NASA Ames planetary scientist Chris McKay. Finally, a more general definition of life – perhaps too general in that it might apply to some nonliving systems – is "a system that possesses the ability (*homeostasis*) of maintaining form and function through feedback processes in the face of changing environments."

What is required to maintain terrestrial life? Many different things are required for different forms of life, but the essentials are organic (carbon-based) molecules for structure and processes, liquid water as an energy and information transporting medium, and a source of usable energy (most often from the Sun, but Earth's heat can be a source as well).

12.1.2 Basic structure of life

All known life-forms live on Earth and are based on the same small set of molecular units and chemical reactions. Four types of essential molecules are *organic* (contain carbon and hydrogen) and account for most biological processes and structures: *proteins*, *nucleic acids*, *carbohydrates*, and *lipids*. Carbohydrates are molecules in which the hydrogen and oxygen atoms form a whole number (that is, 1, 2, 3 . . .) of water molecules. Some classes of carbohydrates (*sugars*) are produced by plants using sunlight as an energy source, and water and carbon dioxide as the raw materials. This process, *photosynthesis*, led to fundamental changes in Earth's atmospheric composition early in its history, as we see in Chapter 17.

The molecules that provide the primary structural material for life, as well as contribute crucially to its functioning, are called *proteins* (from the Greek word *proteios*, or primary, hence "primary substance"). Proteins are long chains (or *polymers*) of relatively small molecular units (*monomers*), called amino acids. An example structure of an amino acid is shown in Figure 12.2. The "R" group distinguishes the particular amino acid – it could be hydrogen or methyl (CH_3) or more complicated combinations of hydrogen, carbon, and oxygen. Of the vast variety of possible amino acids, only about 20 are found to be the building blocks of the major proteins of life.

Long-chain proteins fold into tight bundles, which give rise to the physical and chemical behaviors associated with particular proteins. A typical protein chain may contain from about 50 to 1,000 amino acid molecules strung together. The total number of possible proteins is vastly more than the relatively few (of order 100,000) that actually occur in terrestrial life. Of those that do occur in cells, some play a role in defining the cellular structure, some act to transport or store molecular compounds, and others act as catalysts to control the rates of biochemical reactions; the latter are called *enzymes*.

Proteins cannot make copies of themselves; in the absence of some directive agent or template, the faithful production of

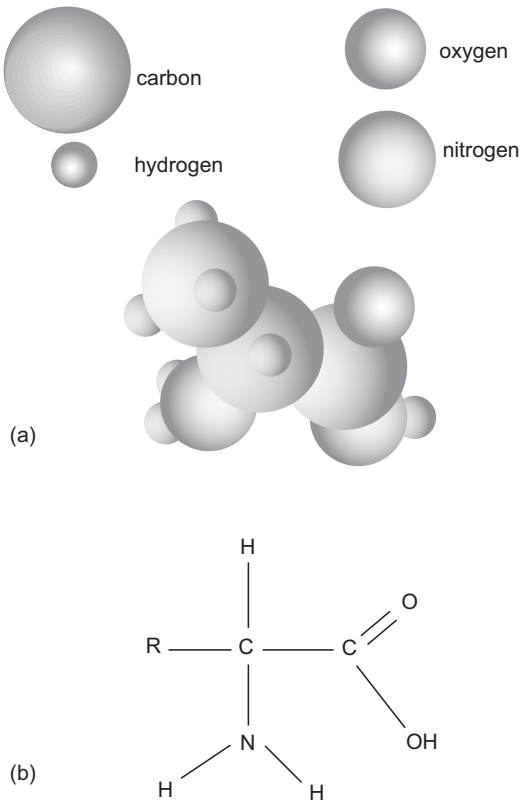


Figure 12.2 (a) Atomic structure of the amino acid alanine, used in proteins. (b) Schematic structure of many amino acids, including most biological ones, where "R" represents a functional group of atoms that defines the particular amino acid.

proteins from the simpler amino acids would not occur in cells. Nucleic acids are molecules that form the building blocks of the templates, which we consider next.

12.1.3 Information exchange and replication

The information-carrying and replicating (or *genetic*) components of terrestrial life are types of nucleic acids called DNA (deoxyribonucleic acid) and RNA (ribonucleic acid). DNA molecules are long double chains normally twisted into a helical structure. The side rails of the double chains consist of a string of alternating sugar and phosphate molecules. Sugar is a simple carbohydrate. Many common sugars, such as glucose, have the chemical formula $C_6H_{12}O_6$; others have slightly different ratios of carbon to hydrogen and oxygen. Phosphate molecules, or phosphate groups, form high-energy bonds in living systems; a phosphate group involves phosphorus and, for example, would have the formula PO_3H_2.

The cross-ties of the DNA chain consist of pairs of four different types of *bases*: adenine (A), thymine (T), guanine (G), and cytosine (C). These bases consist of carbon, nitrogen, oxygen, and hydrogen in complicated ring structures (Figure 12.3). The combination of a base with the sugar and phosphate backbone is called a *nucleotide*.

The pairing of the nucleotide bases is restricted: A with T, and G with C. Thus, the two sides of the chain (*conjugates*) are

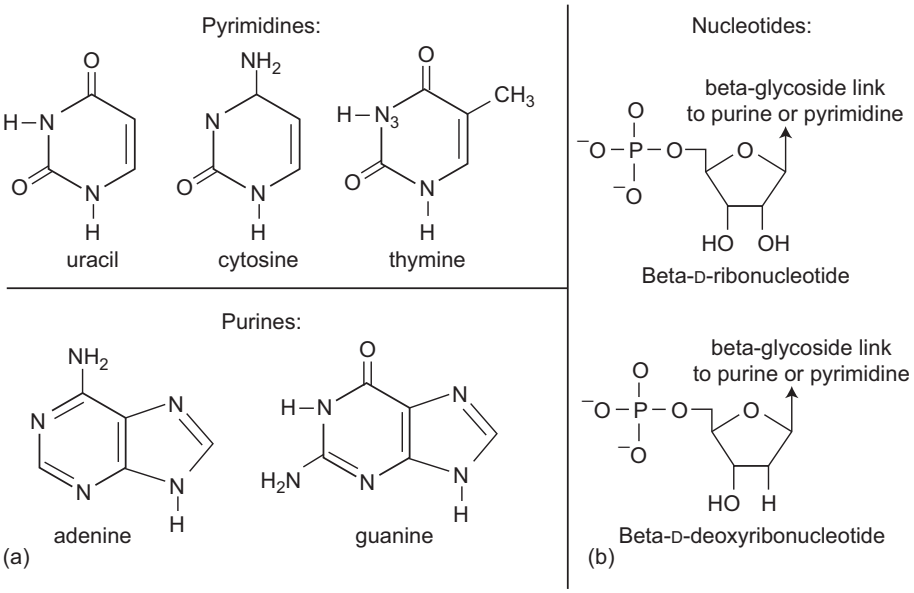


Figure 12.3 (a) The five types of nucleic acid bases in DNA and RNA, showing the characteristic ring structure. (b) Two types of nucleotides are produced from the bases: (top) a ribonucleotide that is the foundation for RNA and (bottom) a deoxyribonucleotide that is the foundation for DNA. Empty vertices correspond to carbon paired with zero or one hydrogen atoms; double lines indicate two shared pairs of electrons. Redrawn from Mason (1991).

redundant to each other because, from the letter on one side, you know what the letter on the other side must be. In replication, the net result is that the two sides of the chain are split, with each side reconstituting (through the mediation of enzymes) its conjugate, resulting in two copies of the original DNA.

12.1.4 Formation of proteins

Protein synthesis is governed by DNA, through the intermediation of RNA. The synthesis begins when DNA, instead of replicating to make new DNA, transcribes RNA. RNA differs from DNA in two aspects: the sugar is of a different form, and the nucleic acid base uracil (U) is present in place of thymine (T). These are relatively minor structural changes in the molecule (Figure 12.3), a fact that we return to in Chapter 13 as we consider the origin of the genetic code.

Thus, a chain of RNA contains a long sequence of molecular monomers chosen from among the four nucleic acid bases A, U, G, C. This chain of monomers can be "read" as a sequence of three-letter "words" constructed from a four-letter "alphabet." Each three-letter word is called a *codon*. Some examples of words are **GUA**, **AAG**, **UGA**. The number of possible words is $4 \times 4 \times 4 = 64$.

Each codon codes for a specific amino acid; thus, the sequence of codons in an RNA molecule (which, remember, is ultimately derived from the sequence in the original DNA) specifies a sequence of amino acids. This amino acid sequence constitutes the synthesized protein. A particular amino acid generally is coded for by more than one codon, because 64 codons are available for the 20 amino acids commonly used in terrestrial biology.

The actual protein synthesis is a bit more complicated, with *messenger RNA* carrying the protein-structure information from the DNA, *transfer RNA* attaching to specific amino acids and aligning them based on the messenger RNA sequence, and *ribosomal RNA* (located in a cellular structure called the ribosome) receiving the ordered amino acid sequence (ferried by the transfer RNA) and acting as a catalyst for final assembly of the amino acid chains. Other RNA molecules assist in DNA replication and in the construction of the messenger RNA. This diverse range of roles for a single kind of molecule makes tempting the proposal that, at some time in the distant past, RNA was central to the genesis of life as we know it. By contrast, DNA, which is not terribly dissimilar to RNA, has a very specialized function as a record of the genetic information of the individual organism and (in separate DNA strands) of certain structures in the cell. This essential but much more limited role compared to that of RNA suggests that DNA is a subsequent, derived molecule.

A length of DNA that carries the genetic information that is ultimately expressed as a single protein is called a *gene*. The genetic code is the complete sequence of nucleotides in DNA, which determines the form and function of an organism's proteins. All living organisms on Earth that have been examined use DNA and RNA to record and express the proteins of which they are made.

12.1.5 Mutation and genetic variation

The replication process sketched above operates with high fidelity. Errors are rare but occur. These errors are called mutations. Such errors, changes in the structure of the DNA, may have a variety of causes such as chemical impurities in the environment or radiation (ultraviolet photons or particle radiation). Other errors or changes may be a result of accidental mixing or crossover of DNA chains in normal cells. Mutations give rise to

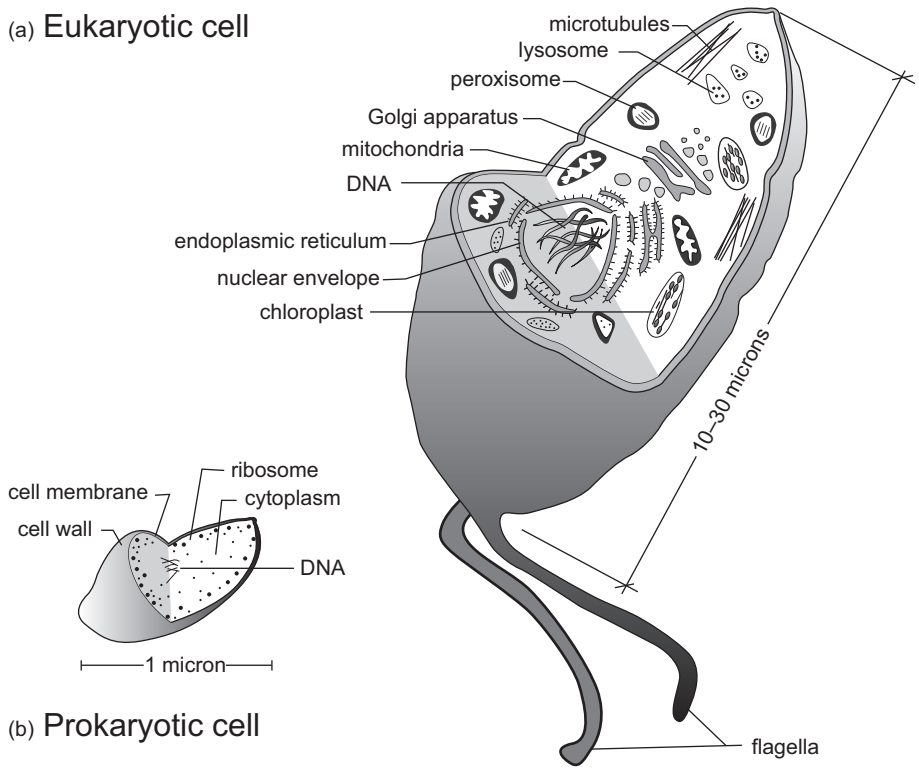


Figure 12.4 (a) Generalized eukaryotic cell, with structures and organelles shown. (b) Prokaryotic cell (a bacterium).

changes in organisms. This genetic variation is usually harmful but sometimes not.

Such variation forms the biochemical basis for the evolution of one species from another, via natural selection within a given environment or through environmental changes in the ecosystem itself. The large-scale pressures for the evolution of species are discussed in Chapter 18, but, without the imperfection and vulnerability in the genetic code that allows changes (both good and bad), such evolution would not be possible, or too slow to be relevant to the history of life on Earth.

Since it is now possible to analyze the genome of an organism and determine the sequence of base pairs, the concept of a molecular clock based on the mutation rate has assumed great importance in estimating when different organisms diverged from one another. The rate of mutation varies from species to species, and even between different components of DNA within a given species – for example the mutation rate of DNA contained in the mitochondria of eukaryotes (see next section) is generally higher than that of the DNA in the nucleus. This molecular clock may have errors of factors of ten or more. In some cases, the mutation rate can be cross-checked with other evidence. For example, the differences in DNA among different peoples can be cross-checked with the migration patterns established by archeology to determine a mutation rate. And in closely related species, such as humans and chimpanzees, it is reasonable to assume mutation rates that are similar, allowing a molecular determination of how long ago the two lineages diverged from a common lineage; to some extent this can be cross-checked by dating fossil remains (Chapter 20).

12.2 The basic unit of living organisms: the cell

With the exception of viruses and virons, which are essentially strands of DNA or RNA sheathed in proteins and which cannot survive independently of other organisms, all Earth life is organized into cells. These structures provide a boundary or membrane for separating the outside environment from the internal one where biochemical reactions occur, and house the DNA and RNA genetic machinery for replicating the particular organism.

Two basic types of cells exist today on Earth (Figure 12.4). *Prokaryotic* (from the Greek "pro" for before and "karyon" for nut, hence seed or nucleus) cells include the common bacterium. The interiors of the prokaryotic cells are dominated by *cytoplasm* (a salt-water medium containing proteins) in which are contained a loop of DNA and *ribosomes* – structures hosting the RNA for protein production. The cell walls are composed of sugars and *peptides*, chains of amino acids that are shorter than the full-fledged proteins. These simple cells are as small as 10^{-6} meters, are able to store energy by either aerobic or anaerobic processes, and divide by a simple splitting process. Most can move by means of attached protein strands called *flagellum*.

Eukaryotes (from the Greek "eu" for good, hence true, and "karyon" for nucleus) are much bigger, more complex cells. The DNA is housed in a *nucleus*, and other structures called *organelles* also occupy the cytoplasmic space. These include *plastids* (the commonly known green ones being *chloroplasts*) in plants, which are the sites of photosynthesis (see section 12.4), and *mitochondria*, within which respiratory processes take place. Most eukaryotes are obliged to utilize oxygen in

their sustaining processes; there are a lesser number of anaerobic eukaryotes. Furthermore, the cells are 10 to 100 times larger in diameter than prokaryotic cells, and reproduce by somewhat more complex processes, which ensure the presence of nucleus and organelles in each new cell.

Clues to the origin of eukaryotic cells lie in the ability of most to take advantage of oxygen and the resemblance of organelles in size and structure to bacteria. They seem to be later arrivals in the history of life, though how late is controversial, with some well preserved organic molecules typical of photosynthesizing eukaryotes dating back to 2.7 billion years.

12.3 Energetic processes that sustain life

Chapter 13 considers life as a phenomenon of nonequilibrium thermodynamics, driven and sustained by substantial flows of energy. Life on Earth primarily utilizes the Sun for energy, with heat from Earth's interior as a secondary source. To appreciate the coupling of life to such energy sources, we must understand how they are utilized by living organisms on Earth.

12.3.1 Common metabolic mechanisms

Energy for living processes requires a usable raw material and suitable chemical reactions to store energy in chemical bonds, which then can be utilized by the organism. *Fermentation* and *respiration* are the most common metabolic processes used by organisms today (Figure 12.5). In each case, energy is stored by the organism in bonds involving the element phosphorus. Molecules such as adenosine triphosphate serve as the storage medium through their phosphate bonds; a single phosphate bond stores 7.3 kilocalories of energy, a large amount. (Biochemists use the unit calorie, but so do nutritional scientists: confusingly, the nutritional "calorie" listed on a cereal box is 1,000 times that of the physicists' calorie, or 1 kilocalorie. Only the physicists' calorie is used in this book.)

In fermentation, which is practiced by bacteria, the sugar, glucose, is split into two molecules of pyruvate. Two phosphate-bonds worth of energy are used to break the bond, but the resulting reaction produces a total of four phosphate-bonds worth of energy. The pyruvate then is converted into ethanol and carbon dioxide, or into lactic acid (depending on the type of bacteria) as waste products. Net energy gain is two phosphate-bonds or 14.6 kilocalories of energy per glucose molecule.

Respiration takes advantage of the presence of free oxygen (O_2) in Earth's atmosphere to extract much more energy from the glucose molecule than fermentation can. Pyruvate again is produced as in fermentation, but instead of immediate conversion of pyruvate into waste products, a complex series of chemical reactions with six oxygen molecules leads to the production of carbon dioxide, water, and 34 additional phosphate-bonds worth of energy. The net result is 36 phosphate-bonds, or 263 kilocalories, worth of energy. A number of biological catalysts, that is, enzymes, are required to mediate and control the *citric acid* cycle that produces the additional 32 phosphate bonds. Although some bacteria do undertake respiration, eukaryotes take the greatest advantage of this process. As we discuss in Chapter 18, the onset of oxygen in Earth's atmosphere was likely the enabling factor for the dominance of multicellular eukaryotic life, with its specialization of cells and high degree of mobility. Confined to only one-eighteenth the amount of energy per glucose molecule, as is the case in fermentation, living processes would be much too sluggish to sustain macroscopic animals.

12.3.2 Photosynthesis

The source of glucose and other sugars used in metabolic processes must lie in an energy-collecting process. Without some means to create such sugar, limitations of food supply for metabolic processes would be far more severe than they actually are. *Photosynthesis* is the production of sugars from water and carbon dioxide, using sunlight as the energy source. Chemically the reaction (in plants) is $6\,CO_2 + 6\,H_2O \rightarrow C_6H_{12}O_6 + 6\,O_2$, where the sugar (glucose) appears as the first compound on the right side of the equation. Energetically, sunlight charges a natural battery in the plant: a molecule called *chlorophyll* is able to donate an electron upon absorption of photons. The source of the electron in plants and most photosynthesizing bacteria is a water molecule. The electrons so liberated then are used to drive the formation of high-energy phosphate bonds, which, in turn, the plant uses to produce sugars.

There are several varieties of chlorophyll and chlorophyll-type molecules utilized by different photosynthesizing organisms. Modern plants employ water as the electron source, and produce molecular oxygen as a waste product. Some bacteria use a less efficient cycle in which the chlorophyll-type molecule is the electron source itself, and the electron then is returned to the donor molecule. This more primitive *cyclic* photosynthesis does not produce molecular oxygen, and captures less energy from a given amount of sunlight than does plant photosynthesis. One type of cyclic photosynthesis, as an example, begins with hydrogen sulfide and ends with sulfur: $2\,H_2S + CO_2 \rightarrow CH_2O + H_2O + 2\,S$.

Some bacteria that conduct cyclic photosynthesis are in fact intolerant of oxygen. Others switch between oxygen-free photosynthesis and respiration. As discussed in Chapter 19, the rather late occurrence of large amounts of molecular oxygen in Earth's atmosphere suggests that the less efficient cyclic photosynthesis dominated early on, and that the oxygen-producing form of photosynthesis was a later innovation.

12.4 Other means of utilizing energy

Respiration and photosynthesis are currently the predominant means of producing storable energy from sunlight and then utilizing that energy for biological processes. However, other mechanisms are employed by organisms that are not in environments where they can photosynthesize or gain access to photosynthesized sugars. *Chemisynthesis* is employed by organisms that live in the environment around deep-sea volcanic vents, where hot, hydrogen sulfide-rich waters pour out of newly formed ocean crust (Figure 12.6). Such waters, compared to the colder, sulfide-poor adjacent regions, have an abundant supply of *free energy*. This term refers to a source of energy that

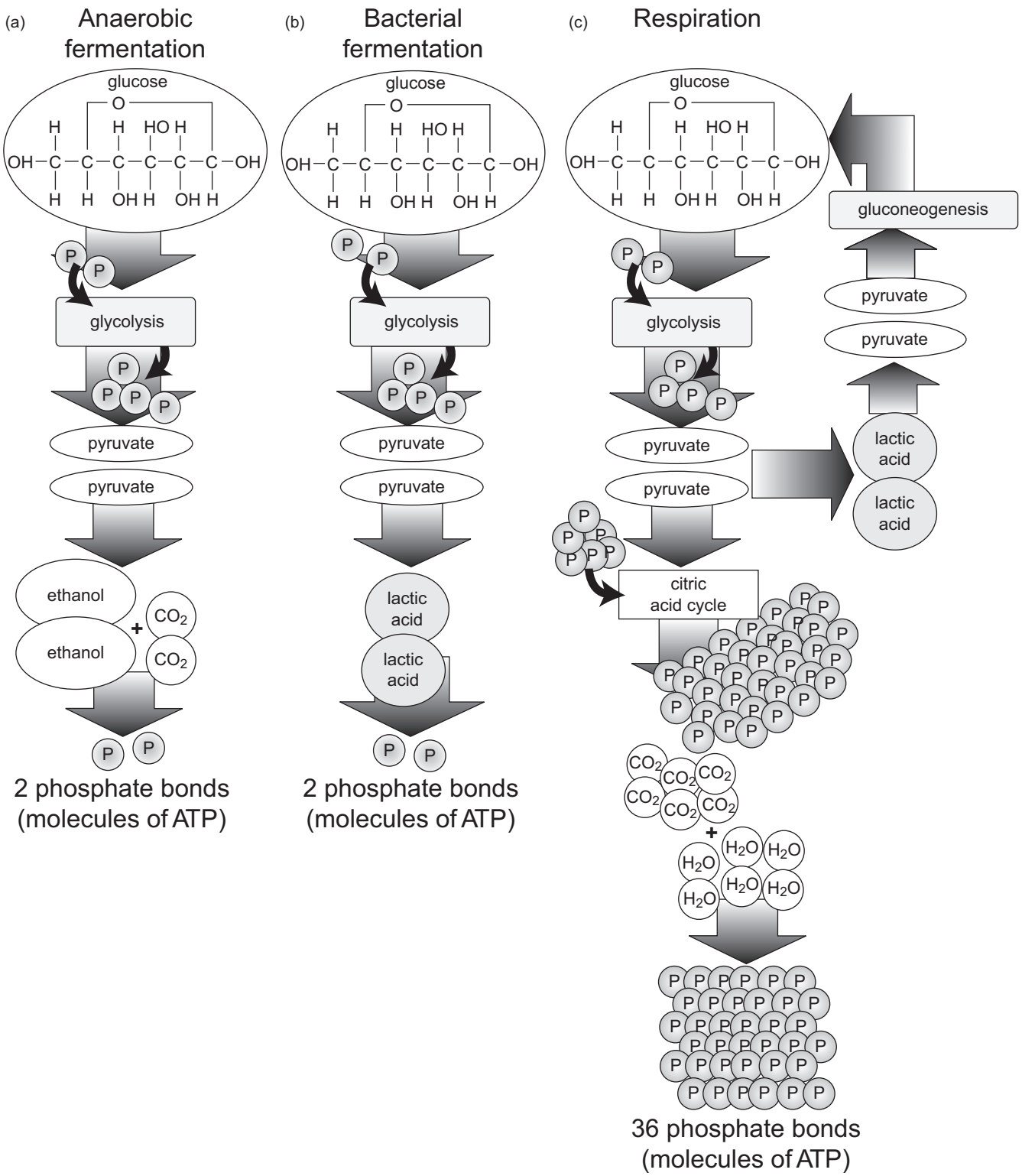


Figure 12.5 Comparison of the common metabolic processes of fermentation and respiration. Various molecules involved in the process are indicated by circular boxes, but only glucose's structure is shown. Square boxes refer to processes; for example gluconeogenesis is the production of glucose. Phosphate bonds are symbolized by a P. Note that fermentation sequesters for the organism, in the end, two phosphate-bonds worth of energy; respiration sequesters 36 such bonds. Modified from Gould and Gould (1989).

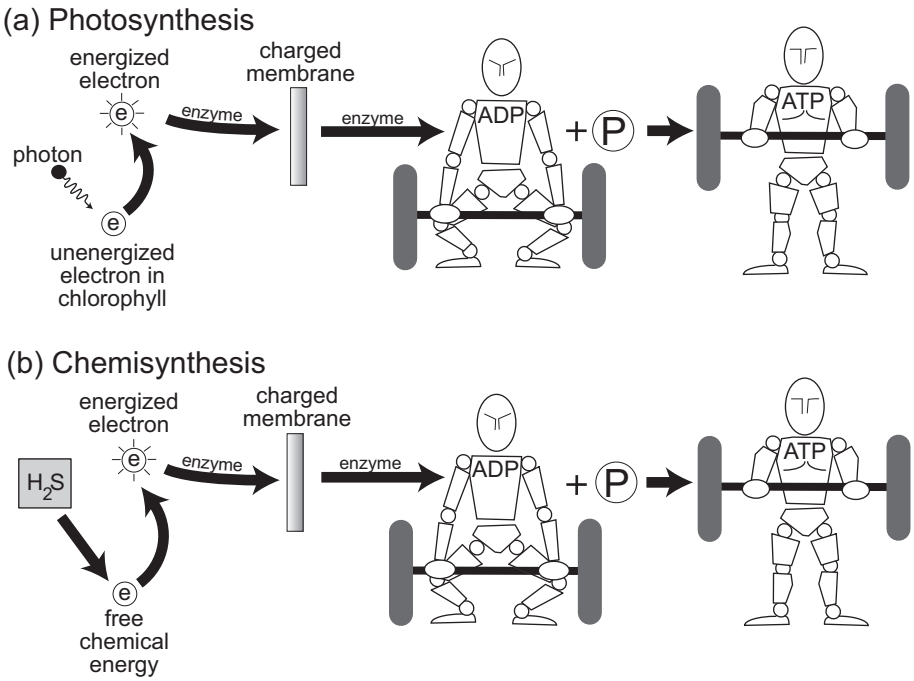


Figure 12.6 Comparison of photosynthesis and chemisynthesis: both allow organisms to sequester useful energy for cellular functions in the form of the phosphate bonds of the molecule adenosine triphosphate (ATP). However, photosynthesis derives energy from sunlight whereas chemisynthesis garners energy from reduced molecules (such as H_2S) not in chemical equilibrium with their surroundings. ATP is symbolized as the energized weightlifter; the preceding molecular form that does not have the energetic phosphate bond is adenosine diphosphate (ADP). Redrawn and modified from Gould and Gould (1989).

can be utilized readily to do some form of work, such as sustain biological processes, or can be stored in high-energy phosphate bonds.

One readily available means to extract energy from the vents is to combine hydrogen sulfide with oxygen to form sulfur dioxide with production of energy. Such a process is possible in an ocean that has free oxygen available, but would not work on the primitive, pre-oxygen-rich Earth. Other biochemical cycles that use sulfur but not oxygen are conducted by some prokaryotic organisms, but these capture much less energy than the oxygen-driven cycles. As with fermentation, chemisynthesis without free oxygen was the hallmark of a rather sluggish primitive biota.

12.5 Elemental necessities of life: a brief examination

12.5.1 Why carbon?

Why is the element carbon the basis for biochemistry? Of all elements, carbon possesses the greatest tendency to form covalent bonds and, in particular, has a remarkable tendency to bond with itself. These characteristics reside to some extent with all elements near the center of the periodic table (Figure 2.6), for which there is not a strong tendency to favor donation over acquistion of electrons. However, carbon is most distinguished in this regard as a small element (few electron shells) and being positioned in the central column IVA of the table. It readily forms a variety of long-chain, sheet, and ring structures, many of which play important roles in the basic biological molecules, as seen in Figures 12.2 and 12.3. In addition, carbon is the fourth most abundant element in the cosmos (after hydrogen, helium, and oxygen), being made readily from helium fusion in stars.

The element silicon has chemical-bonding properties very similar to those of carbon, being one row below the latter in the periodic table. Silicon also forms chain, sheet, and ring structures with nearly (but not quite) the same ease and variety as does carbon. It is not nearly as abundant as carbon in the cosmos because it is the product of a later stage of fusion reactions achieved only in massive stars, but, after oxygen, it is the primary constituent of the crust of Earth. The similarities in silicon's properties to those of carbon and its high abundance in geologic materials are responsible for the lithification of biological remains in the form of fossils, as discussed in Chapter 8. It is natural to ask whether silicon might be the elemental basis for biology on another planet in another planetary system, perhaps one on which conditions are not quite suitable for carbon-based life because of higher temperatures, or where carbon-bearing molecules were not supplied to the planet's surface, for whatever reason.

The biologist A. Cairns-Smith has proposed that, on the early Earth, layers of crystalline silicates might have served as the basis for a very primitive kind of life, which served in turn as a sort of template for, and gradually evolved into, carbon-based life. Certainly, some clay minerals have a surface upon which organic molecules can adhere, in favorable positions, which

might have encouraged the formation of replicative molecules similar to RNA (or perhaps RNA itself). Cairns-Smith's proposal is more radical, however, and represents a kind of inverse fossilization in which the end result is a vigorous and robust carbon biochemistry.

Because any evidence for a putative primitive silicon-based life surely has vanished, such speculations must remain just that. However, consider that, in natural chemical systems, very small intrinsic advantages can be magnified into very large effects. In addition to forming stronger bonds with itself (C–C) than can silicon, carbon is more versatile in its bonding properties, and hence can form a wider variety of structures. The *information* content, therefore, inherent in carbon-based biochemistry is larger than that available using any other single element as a basis. The American chemists J. Feinberg and R. Shapiro suggest the analogy of an alphabet. By way of example, the English alphabet has 26 characters, from which are made words and sentences. Other human languages may have hundreds or thousands of symbols. A computer, on the other hand, uses two basic states or characters (usually symbolized as 0 and 1) to carry information. A computer can record the same information in its two-letter alphabet as an English writer would in his or her 26-letter code, but the computer must use longer words and longer sentences to record the same concept or amount of information.

Biochemistry in which the number of possible compounds is fewer than in a carbon-based system, by analogy, would have to utilize those fewer compounds to build sufficient functional structures and information-carrying molecules to survive and perpetuate its particular kind of chemistry. Organisms based on such a biochemistry might build essential structures more slowly and less efficiently, reproduce less frequently, and perhaps be limited to more narrowly characterized environments than is carbon-based life. In 1961 Carl Sagan noted that silicon cannot form the changeable and versatile side chains that are crucial to the proteins and nucleotides that characterize carbon-based life. Thus there might well be some environments that favor silicon over carbon but, in a general sense, carbon-based biochemistry would more likely be fruitful and multiply, gain access to and utilize a variety of sources of free energy in its environment, and respond robustly to drastic environmental changes, than would its silicon-based counterpart.

There is one environment, however, in which silicon-based life is in a sense extant on Earth today. The term *artificial life* has been co-opted by information scientists and physicists to refer to computer programs that move bits of information around within the silicon-based processors and memories of modern computers and, in doing so, simulate the biological processes of complexification, growth, and reproduction. The significance of such programs is highly controversial, and most laboratory biologists do not regard the resulting transfers of energy and information within the computers as equivalent to living processes. We return to artificial life and its lessons for terrestrial carbon-based biology in Chapter 13.

Carbon–silicon is not the only elemental substitution under discussion. Phosphorus, so essential to life through the phosphate bond in DNA, RNA, ATP, and ADP, sits one row above arsenic in the same column. Might arsenic substitute for phosphorus? In 2010, NASA scientists claimed they had found a bug that could tolerate eating arsenic in place of phosphorus in arsenic-rich conditions, but it has yet to be established whether the arsenic is really substituting for the phosphorus in the biomolecules. Should it turn out to be the case, the implications for potential exotic biochemistries elsewhere in the cosmos would be enhanced. On the other hand, should arsenic merely be adhering to a DNA or RNA molecule as a kind of contaminant, it would confirm the longstanding objection to use of arsenic in terrestrial-type life, namely the very rapid breakdown of the arsenic equivalent of phosphates in the presence of water.

12.5.2 Why water?

The third most abundant element in the cosmos is oxygen, and it occurs primarily in two molecular forms: carbon monoxide (CO) and water (H_2O). In our own solar system, the former is rare, presumably because its high volatility discouraged the trapping of large amounts in solid material. Water ice, on the other hand, was a primary "mineral" in the primordial solar system, and Earth may have had access to large amounts of it through comets. Water is thus abundant on Earth and elsewhere in the solar system, but its importance for life is not due solely to abundance: the properties of water are crucial for the one form of life that we know of, namely our own.

Water exists as a liquid over a temperature range particularly suited for organic reactions – not so cold that reaction rates are too slow to sustain biological processes, and not so hot that organic bonds are too readily broken. Water's existence as a liquid from 273 K up through the critical point at 647 K (beyond which only a single, gaseous/fluid state is stable) is unusually broad and is not significantly overlapped by many other abundant molecular species. (Liquid water in an open container is not stable in our atmosphere above 373 K – the boiling point – because the vapor pressure of water exceeds the ambient atmospheric pressure at sea level. For liquid water to exist above that temperature on Earth, it must be confined to a pressure vessel, such as a pressure cooker.)

This important property of water is due mostly to the somewhat unusual bonding mechanism between water molecules, in which the hydrogen atoms from one water molecule form bonds of modest strength with those of another water molecule. This mechanism of *hydrogen bonding* produces much stronger bonds than those seen in similar-sized molecules such as methane and carbon dioxide, for which the liquid phase occurs at much lower temperature, and much weaker bonds than in silicates, which melt at much higher temperatures. The hydrogen bond itself, caused by small residual positive charges at the hydrogen end and negative charge at the oxygen end of each molecule, is also responsible for making liquid water "polar", that is, a good conductor of electricity and a good solvent for materials of biological importance.

Water functions as the medium within which biochemical processes take place, and by virtue of being a liquid, nutrients and wastes can be transported in a controlled fashion. Cellular membranes are important boundaries within which the salinity and acidity of the water are carefully regulated by biochemical processes, so that nutrients can be properly absorbed through the membrane and waste products excreted. In complex organisms with circulatory systems, blood is essentially a liquid-water

medium packed with cells of various kinds and with a multitude of functions; the liquid nature of the blood enables transport over large distances to be rapid and efficient.

Most biologists would argue that, of all the requirements of life, the need for liquid water or a similar liquid medium is paramount. In its absence, the ability to selectively pass nutrients into loci of biochemical activities (cells, in the terrestrial case) and to remove undesired products of biochemical reactions is extraordinarily limited.

It has been suggested that other liquids, such as ammonia–water solutions or hydrocarbon liquids, could be substitutes in low-temperature environments. The properties of ammonia–water solutions are such that certain kinds of organic reactions on which terrestrial life depends are prohibited; Shapiro and Feinberg suggest that weaker chemical bonds, such as those involving nitrogen, would tend to be more important. Such biota might be less robust than terrestrial life, and if the ammonia–water solution were near its freezing point (as low as 176 K), organic reactions would be much more sluggish than at room temperature.

Methane and ethane, stable at temperatures appropriate to Saturn's moon Titan, have properties that are even more alien to Earthly life than ammonia–water: in the absence of other molecules these hydrocarbon liquids would be completely nonpolar, eliminating the usual electrochemical processes that enable molecules to fold and transport species across cell walls. However, as suggested by Steven Benner and colleagues, a nonpolar hydrocarbon liquid, because it cannot hydrogen bond, might allow other organic species to themselves combine through hydrogen bonding – a novel kind of chemistry that is blocked in liquid water. Further, in the absence of oxygen-bearing species in such hydrocarbons – the situation we find on Titan – carbon–nitrogen bonding might substitute for at least some of the roles that carbon–oxygen bonds play on the Earth. Would these novelties permit an exotic kind of life to exist that uses hydrocarbon liquids in place of water? The great complexity of organic chemistry, made possible by the flexibility of carbon discussed in Chapter 2, renders it infeasible to decide one way or the other in the absence of sampling of a natural "lake" or "sea" of liquid hydrocarbons. The possibility of doing so on Titan is discussed later in the chapter.

12.5.3 Is free oxygen essential?

One common misconception is that life on Earth requires molecular oxygen (O_2) for its existence and propagation. In fact, as discussed above, molecular oxygen and associated respiration processes are only one kind of energy-producing mechanism that sustains life. Many simple organisms not only do not use oxygen in metabolic processes, but are poisoned by it. Because molecular oxygen reacts readily with many organic compounds and breaks bonds in such compounds, organisms that tolerate the large amounts of free oxygen in Earth's atmosphere had to evolve protective mechanisms to avoid undesirable oxidation of organic molecules. Additionally, the presence of free oxygen made utilization of nitrogen as a nutrient more difficult. However, the very large amount of energy-storing phosphate bonds that can be created using molecular oxygen outweighed the disadvantages, and complex cells and multicellular life became possible as O_2 abundance in the atmosphere increased over time (see Chapters 17 and 18).

Strategies for searching for inhabited planets around other stars focus on very sensitive spectroscopy to detect the presence of molecular oxygen in the atmosphere. Remember, however, that a negative result does not indicate the absence of life, but rather indicates either the absence of planet-wide photosynthetic processes that produce large amounts of molecular oxygen, or the existence of a sink to effectively soak up the oxygen that is produced. Life utilizing anaerobic processes, such as fermenting bacteria or the sulfur metabolizers at deep-sea vents, might still be present. This does not mean such strategies are flawed; the detection of molecular oxygen is feasible with large interferometers planned for later this century, and hence a worthy goal particularly if one is interested in finding advanced life. However, there might well be a multitude of planets throughout the cosmos on which life exists but, for whatever reason, remains in a stage in which less efficient anaerobic metabolisms predominate.

12.6 Solar system sites for life

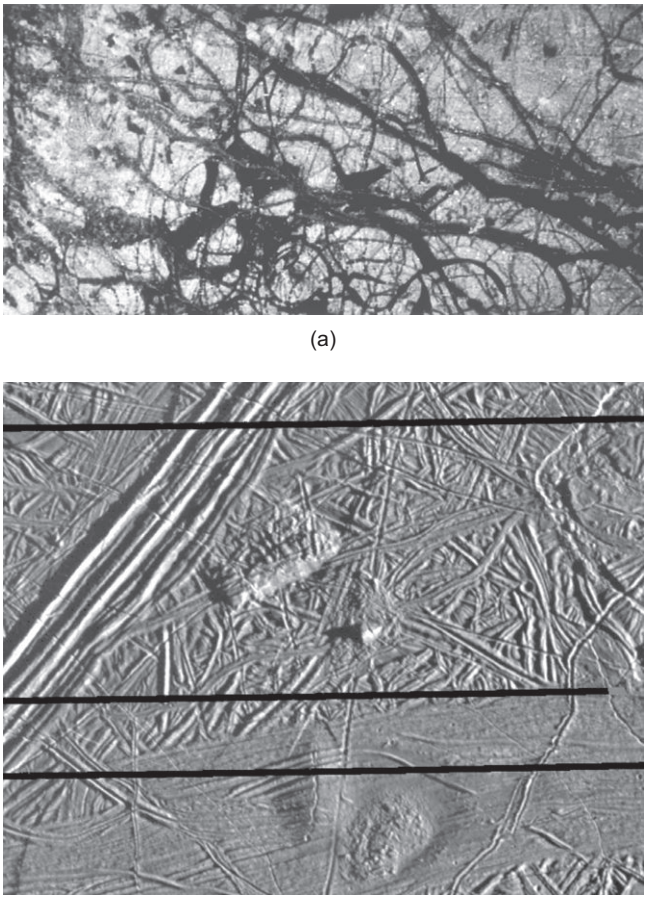
Although liquid media other than water are possible sites for life, a well-constrained search for solar system environments in which life might have arisen should focus on liquid water – because we know that life arose at least once in such a medium! In our own solar system, there are four environments within which liquid water is known to be stable at or near the surface, or was likely stable for long periods in the past: Earth, Mars, the interior of Europa, and the water clouds of the giant planets. Saturn's moon, Titan, although too cold for liquid water to be stable, is rich in organic molecules and has vast lakes and seas of liquid hydrocarbons. Also, large impacts on Titan may have provided energy to melt its ice crust and provide liquid water for relatively short periods of time.

12.6.1 Atmospheres of the giant planets

The atmospheres of the giant planets represent the most speculative site for life, one proposed by Carl Sagan and others some decades ago. As discussed in Chapter 11, no solid surface exists except at enormous depths in the interiors of these gaseous bodies. The water clouds lie below a layer of ammonia clouds, which in turn lie below methane clouds in the cases of cold Uranus and Neptune. Living organisms there might be composed of structures that allow them to cycle in depth from the relatively warm water clouds up to the ammonia clouds, absorbing sunlight at the higher levels and various trace organic constituents at a variety of altitudes. Although one cannot rule out such a biota, the initial evolution of such organisms to a point of sophistication such that they could safely cycle in altitude without sinking into excessively hot depths is an open issue.

12.6.2 Interior of Europa

The shallow interior of Europa, a satellite of Jupiter just 10% smaller than our own Moon, contains liquid water and hence is a potential venue for life. Spectroscopic studies of its surface



(a)

(b)

Figure 12.7 *Galileo* views of Europa. (a) Large-scale view. (b) Close-up image showing details around one of the crack systems. Courtesy of NASA Jet Propulsion Laboratory.

from Earth reveal that water ice is an important or predominant component. Radar signals bounced off of Europa from Earth are reflected back with very high intensity, like a mirror, again indicating the predominance of water ice. Photographs taken by *Voyager 2* in 1979 reveal a bright surface covered in cracks, which themselves are only slightly darker than the surrounding surface. So bright is Europa's surface that most of the sunlight hitting it is reflected, and the average surface temperature is below 110 K, colder than the darker surfaces of Ganymede and Callisto.

From 1996 to 2003, the sensitive electronic camera aboard the *Galileo* Jupiter orbiter imaged the surface of Europa in much greater detail than could *Voyager*. Preliminary study of the images yields additional circumstantial evidence that a liquid water layer exists beneath the ice. These include an enormous variety of cracks of different degrees of freshness, areas where cracks have been cut or buried by flows of liquid or warm ice, pieces of crust that have tilted upward as if floating on a layer of liquid water beneath, and craters in the ice displaying softened edges consistent with a thin ice crust (Figure 12.7).

The magnetometer aboard the *Galileo* orbiter recorded signatures of Jupiter's magnetic field as it passed by each of the large moons of Jupiter, and provided the best evidence for a subsurface liquid water ocean on Europa. In contrast to Ganymede, which possesses its own magnetic field, Europa simply distorted the shape of Jupiter's field as it moved through it on its orbit. The amount of distortion was strong and easily measurable from different angles on different flybys of the *Galileo* orbiter, and could best be explained if an electrically conducting layer existed inside Europa but close to its surface. To fit the *Galileo* data, the layer must be so electrically conducting that only a salty, liquid water ocean is plausible (one could invent other possibilities like molten metals and so forth, but these could be neither close to the surface nor molten.)

Europa's surface layers may be cracked water ice at very low temperatures, but the density of this moon tells an intriguing story about the interior that is also important to the possibility of life there. At 3.0 g/cm^3, Europa cannot be composed entirely of water ice, which has a density close to 1 g/cm^3 (varying somewhat with pressure). To match the density with water ice and rock requires a moon composed of 80% rock and only 20% ice by mass. The rocky component has embedded within it the radioactive isotopes of potassium, uranium, and thorium and, as described in Chapter 10 for Earth, the decay of such isotopes produces heat.

Add to this source of heat one other: tidal heating. Both Io and Europa have orbits that are slightly noncircular and are maintained that way by the mutual gravitational pulls of Io, Europa, and Ganymede against each other. These pulling motions are effective because the orbital periods of the three satellites are simple multiples of each other – the period of Europa is twice that of Io, and that of Ganymede twice Europa's. So, like a child on a swing pumping his or her legs in synchroneity with the period of the swing, these satellites tug gravitationally on each other and keep their orbits noncircular. This, in turn, means that even though each moon keeps one face approximately toward Jupiter all the time – as does our Moon to Earth – there is a small amount of twist as each moon varies in its orbital speed. The twisting is enough to cause frictional rubbing of rocks against each other in Io, leading to extraordinary heating that has melted its rocky interior and produced spectacular volcanic eruptions viewed by *Voyager*. Europa, 50% farther from Jupiter than Io is, suffers some tidal heating, but much less than Io experiences.

The preponderance of silicates in Europa is important to the possibility of life not only for the heating they provide, but also for access to elements that tend to be present in silicates and are important to life. Models of the interior of Europa indicate that, almost certainly, the subsurface ocean is in direct contact with the silicates beneath; that is, the base of the ocean is rocky as is the case for the other. Leaching of phosphorus, magnesium, and other elements important for life, as well as the potential maintenance of chemical gradients providing a source of available energy for life, may depend on this property. In contrast, while other giant moons like Ganymede's Titan and even Callisto may have deep-seated oceans, the base of these oceans is a thick stratum of high pressure water ice perched above the silicate core.

Europa, then, almost certainly has a liquid water ocean lying beneath a frigid surface, and a source of heating that maintains the liquid state. What is uncertain is whether Europa acquired

enough carbon-bearing, nitrogen-bearing, and other compounds during its formation to allow for carbon-based life in the ocean. It is also not known just how thick is the intervening ice crust. To determine these important characteristics for life will require a mission to orbit Europa, equipped with radar to probe the ice, precision laser altimeters to measure the shape of Europa as Jupiter distorts it along its orbit, and spectrometers to search for organic molecules that may have seeped to the surface along fractures. Such a mission is challenging because of the intense radiation associated with Jupiter's magnetic field, radiation which bombards spacecraft electronics and sterilizes very quickly the surface of Europa itself.

12.6.3 Titan

Saturn's largest moon, Titan, is bigger than the planet Mercury. It has an atmosphere that has a surface pressure of 1.5 atmospheres and is mostly nitrogen. Methane is the next most abundant gas in the atmosphere. Titan is so far from the Sun that the surface temperature is 95 K. This is so cold that methane and similar molecules exist on Titan's surface as rivers and seas. Most of what we know about Titan comes from the *Cassini–Huygens* mission, which dropped a lander to the surface in 2005 and continues to observe Titan from a complex spacecraft that orbits Saturn and makes repeated flybys of the giant moon (Figure 12.8). Titan's surface and atmosphere appear to be an intriguing mimic of Earth, with methane substituted for water. The action of solar ultraviolet photons creates a complex organic chemistry in Titan's atmosphere, forming a globe-encircling haze that obscured the surface, requiring instruments such as radar and infrared cameras to probe the surface. Some products of the methane chemistry are stable as liquids on Titan's surface, mixing with methane to make the polar lakes and seas. Others are solids and both coat the surface over large areas as well as form fields of dunes in the equatorial regions.

Titan is a complex, planet-sized chemical factory that almost assuredly has no life like that of the Earth because of the low temperatures. However, arguments have been made that Titan is in fact a good place to go to test whether life, in its most generally defined way, is a natural outcome of the availability of suitable molecules, liquids, and sources of energy. Should a form of organized, self-replicating chemistry, based on organic molecules but with methane rather than water as a liquid, be found in the lakes and seas of Titan, it would be a momentous discovery. Exploration of this world is therefore of high interest. However, because much of the chemistry going on in the atmosphere and on the surface may give us insight into organic chemistry on the earliest Earth, Titan is an important place to explore even in the absence of an exotic form of life. Although Titan's atmosphere is probably chemically much more reducing (that is, rich in hydrogen) than the early Earth's, it is likely to be a better analogue to the Hadean Earth than is our present, oxygen-rich, biologically dominated planet. Occasional impacts or volcanic eruptions might melt the crust and produce pools of liquid water and ammonia for some thousands of years or more; within these transient pools interesting prebiotic chemistry akin to that which occurred on the Earth might occur.

12.6.4 The Mars of today and yesterday

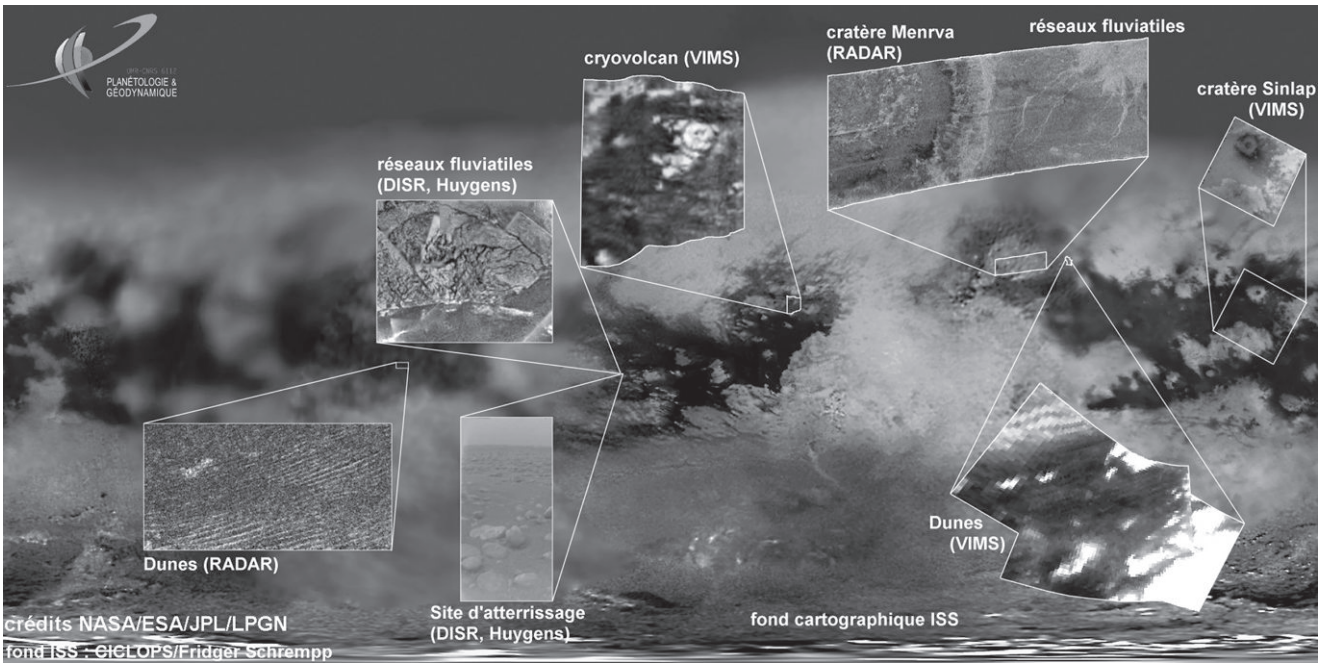
The American *Viking 1* and *Viking 2* robotic landers operated on the surface of Mars beginning in 1976 and extending into the early 1980s. Among the experiments were four designed to determine whether life existed in the upper few centimeters of soil collected by each lander. Each of the experiments was designed to test for a particular kind of energy-generating mechanism, such as photosynthesis, fermentation, and chemisynthesis. Soil was activated through the addition of potential nutrients, and then the experiments monitored release of gases that might indicate biological activity.

Some of the experiments showed positive results, but in combination, the experiments weighed against biological activity in the Martian soils. The positive activity was best explained as reaction of a nonbiological, highly oxidized component in the soil reacting with the gases and liquid nutrients supplied by the experiments. The death-knell for a biological explanation came from the *gas chromatograph/mass spectrometer* experiment, a device that can detect very small amounts of organic molecules. The amount of organic molecules at the two *Viking* landing sites, in the soil, was less than one part per million, and probably at the part-per-billion level for more complex organics. With so few organics, carbon-based life evidently was not present, and speculation about, for example, silicon life was not profitable because no evidence for such life could be obtained from the experiments.

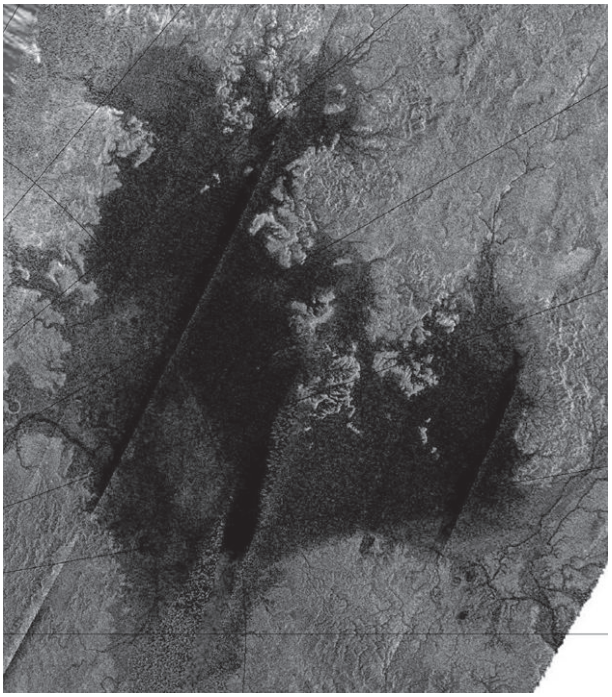
The two *Viking* lander sites, on high plains, were particularly poor candidates for sustaining living organisms, and areas near canyon bottoms, where the thin atmosphere has higher pressure, might be better. Nonetheless, current Martian conditions – an atmosphere of carbon dioxide with a total pressure less than a hundredth that at the surface of Earth, and surface temperatures generally well below the water freezing point – are not promising for life. The search for life in protected enclaves will be extremely challenging by robotic means, and piloted expeditions are decades away. Hence, the focus of research into life on Mars has shifted to finding evidence of past life.

Evidence for ancient clement conditions on Mars is described in Chapter 15, and consist of now-dry valley networks and outflow channels that appear to have been carved by liquid water (or debris carried by water) and more-controversial evidence in the form of possible glacial features and dried lake beds. Clearly something different happened on Mars in the past than today, and ancient conditions appear more promising than those at present. What further buoys the hopes of searchers for evidence of past Martian life is the presence of living organisms on Earth in extraordinary places. Submarine hot springs at mid-ocean ridges are, in the absence of sunlight, a rich abode of life. The dry valleys of Antarctica sport lakes whose surfaces are frozen over year-round, but in which a variety of microbial organisms thrive. Bacteria have been found living in rocks kilometers under the surface of Earth. All these sites could plausibly have had their analogue on early Mars, and in Chapter 15, we place possible life in the context of the history of Mars and the times when such environments might have existed.

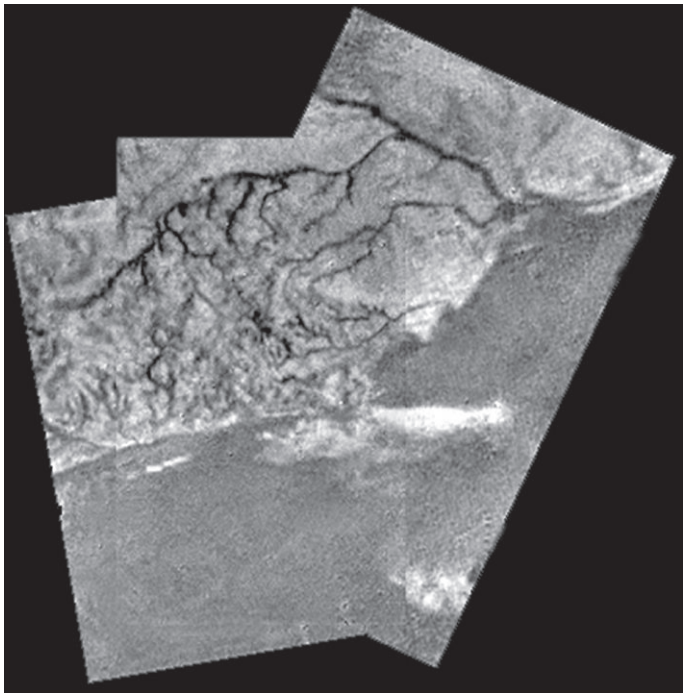
To conduct the search for fossil evidence of life on Mars will be a daunting challenge, because such life very likely was microbial and not large, complex eukaryotic organisms (see



(a)



(b)



(c)

Figure 12.8 *Cassini-Huygens* views of Titan (a) Large-scale view, showing details of surface features from the radar and the VIMS, a near-infrared mapping camera. (b) Radar image of one of Titan's large hydrocarbon seas, called Ligeia Mare, about 800 km across. (c) Image from the Huygens probe under the parachute of a system of dendritic channels stretching over 20 square kilometers; features as small as 30 meters are visible. Images courtesy NASA/ESA/JPL/Space Science Institute/Univ. Arizona. See color version in plates section.

Chapters 15 and 18). However, even microbial life leaves behind signatures. The isotopic ratio of carbon-12 to carbon-13 in organic molecules may be increased from the baseline value by cycling through biological systems, and hence looking for evidence of unusual isotopic patterns in trapped gases in Martian rocks is one avenue. Researchers at the University of California have found that relatively primitive life-forms can have an effect on the kinds of mineral structures that precipitate from seawater on Earth, and the same effect might be preserved in rocks on Mars. Finally, the macroscopic evidence of earliest

abundant Earth life is the lithified remains of bacterial colonies, the stromatolites of Chapter 10. Because evidence for such organisms stretches back to the oldest fossil record, examination of sedimentary Martian rocks for such patterns also should be undertaken.

In August 1996, NASA scientists David McKay and colleagues brought the search for Martian life to the attention of millions of people around the world with an astonishing assertion: that they found evidence for relict biological activity in a meteorite thought to have come from Mars. Meteorite ALH84001, found in Antarctica in 1984, is one of 12 SNC meteorites thought to have been from material blown off of Mars by an impact, and onto a collision course with Earth. These samples contain trapped gases, which, in their chemical and isotopic signatures, are identical to the present atmosphere of Mars as sampled by the *Viking* mission. Work over the past several decades by a number of researchers has shown that it is plausible for a large impact to gouge out a portion of the Martian crust and send some of it toward Earth.

ALH84001 is an old igneous rock, with a radioisotopic age of 4.5 billion years. It is therefore from some of the earliest Martian crust. The meteorite was heated again about 4.0 billion years ago by a strong shock, possibly a nearby impact. Globules of carbon-bearing minerals called *carbonates*, described in more detail in Chapter 14, were found in the rock. These formed later than the rock itself and are tentative evidence for the presence of liquid water flowing through the rock, tentative because carbonates can under some circumstances form in the absence of liquid water.

The age of the carbonate formation in the rock is highly uncertain, with different groups estimating ages from 3.6 billion years (from potassium–argon dating) to 1.4 billion years (from rubidium–strontium isotopes). In either case, these ages are much older than the time when the rock was blown off of Mars. This time is estimated by examining tracks made by cosmic rays on the surface of the rock exposed to space. The abundance of unusual isotopes of noble gases made by cosmic-ray collisions give a residence time in space of between 10 million and 20 million years. Once on Earth, isotopes such as carbon-14 and others of boron and chlorine, also made by cosmic-ray hits, begin to decay; their abundances indicate that ALH84001 has been on Earth only 13,000 years. So, the impact and delivery of this rock to Earth were much more recent than the formation of the carbonates, indicating that the carbonates were formed when the rock was in the Martian crust. Supporting this is the fact that the organic (carbon-bearing) content of the meteorite increases toward the center of the rock, suggesting that at least some of the carbon-bearing material is from Mars.

McKay and colleagues went a step further to propose three lines of evidence that are consistent with biological activity. First, ring-shaped carbon-bearing molecules called *polycyclic aromatic hydrocarbons*, or PAHs, were found near the carbonate globules. Although PAHs can be formed in nonbiological environments, including interstellar space, from which their spectra are detected by sensitive Earth-based telescopes, the structure of the Martian PAH molecules differs from those seen to date from nonbiological sources. However, McKay and colleagues have not shown that biological activity necessarily forms such PAH structures either, and so, by themselves, the PAHs do not argue strongly for biological activity.

The second line of evidence comes from the presence of microscopic crystals of magnetite (Fe_3O_4) in the meteorite. Crystals of similar chemical composition and size are made by so-called "magnetotactic" bacteria on the Earth that can orient themselves according to the direction of the Earth's magnetic field. Such crystals, it was claimed, are not normally formed by abiotic processes expected in the Martian environment; however, a series of experiments performed at NASA's Johnson Space Center showed that impact processes could product magnetite crystals of the right size and composition. The mineral siderite, an iron carbonate with the formula $FeCO_3$, will decompose into magnetite and carbon dioxide when heated, either by the shock of impact or other processes. It therefore seems that the presence of such crystals does not require a biological explanation.

Most controversial are images, constructed by bombarding electrons into the sample, of very small structures near the globules – about 10 to 100 times smaller than terrestrial bacterial. They look like microbial forms, and McKay and colleagues argue that they could be evidence of life. Biologists and geologists today argue about the possible existence of simple cells some 10 times smaller than bacteria that are indirectly inferred to be present in Earth rocks. These *nannobacteria* are a speculative and as yet unproven form of terrestrial life, and their invocation as support for the Martian microbe interpretation has raised more controversy. To show that the forms in ALH84001 are cells requires finding cell walls, which as yet cannot be seen in the images.

The reader should be skeptical of the interpretation of the ALH84001 data in terms of biology. McKay and colleagues argue that the timescales are right – if the carbonates are 3.6 billion years old, then the evidence in the *Viking* images of a more clement Mars at the time are consistent with life beginning then. But the carbonates could be younger.

Further evidence against Martian life in ALH84001 came in 1998 when scientists from the University of California and University of Arizona analyzed the "organic" carbon (that is, the carbon not in the carbonates) from ALH84001 in more detail. From several lines of investigation, including measuring the carbon-14 to carbon-12 ratio, they concluded that most or all of this carbon is terrestrial, not Martian – though they confirm that most of the carbonate phase is likely from Mars. The ambiguities of interpretation associated with the claim of evidence for life in ALH84001 illustrate the great challenge of finding evidence for microbial life from subtle chemical clues and images at very small scales. It will be even more difficult to perform such searches on the surface of Mars; perhaps the best strategy is to locate promising sites on Mars where life may have existed, and then return samples back to Earth.

More recently the discovery of methane in the Martian atmosphere has rekindled discussion over the possible existence, not of fossilized life, but of extant life within the crust of Mars. Detections by an instrument aboard the European *Mars Express* orbiting spacecraft built by V. Formisano of INAF in Italy, followed by ground-based observations by M. Mumma and colleagues in the US, which documented sources in several places on Mars (Figure 12.9), firmly established the existence of this simplest organic molecule. In a CO_2-rich atmosphere bathed in ultraviolet radiation, overlying a reactive surface in which enough free oxygen is available to oxidize quickly any organic

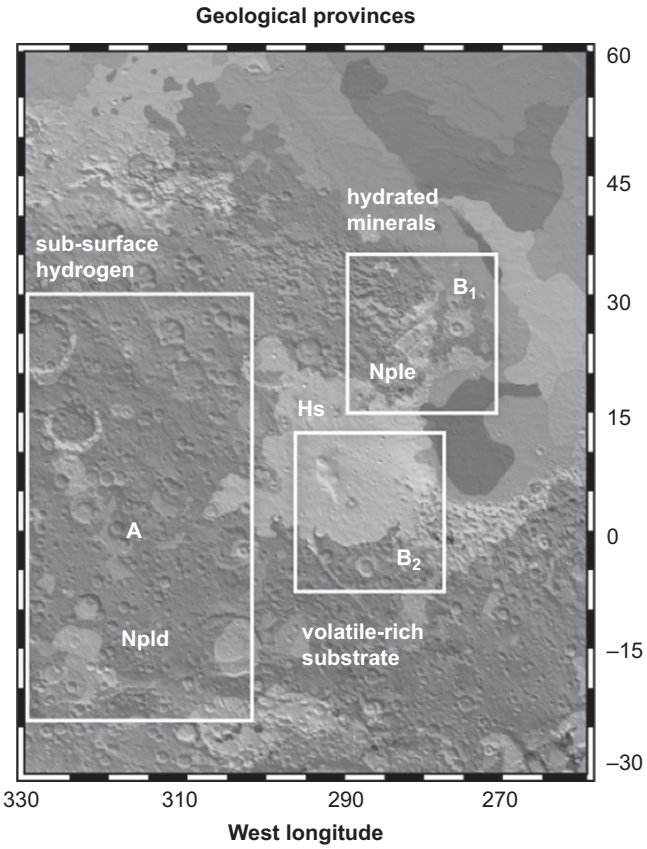


Figure 12.9 Map of Mars showing concentrations of methane gas in a broad region stretching from mid-southern to mid-northern latitudes, and over 60 degrees of longitude. Regions A, B_1 and B_2 are places where the methane emission is localized. Npld and Nple refer to ancient plains estimated to be from the first 20% of Martian history, while Hs, the smooth intervening region, refers to possible volcanic deposits that could be as young as 3.1 billion years or as old as 3.6 billion years. Figure from a color image courtesy of M. J. Mumma/NASA, which originally appeared in Mumma *et al.* (2009).

compound, the persistent presence of CH_4 would seem highly unlikely. Indeed, its abundance seems to have declined over the past few years as if it were the result of a sudden outburst from the crust, followed by destruction in the atmosphere or on the surface. There is no way to know if the methane is a product of biological processes, or instead might be the results of reaction of carbon dioxide with rock in the presence of liquid water. Even the latter, though not a direct sign of life, would portent well for a potentially habitable environment beneath the Martian surface. Deploying a rover or rovers to the general vicinity of the emissions would allow for direct sampling of the gases if the vent could be found. This, however, is technically extremely challenging. More feasible is an orbiter sensitive enough not only to sniff for the methane gas in the atmosphere at the part per billion level, but perhaps also to measure the isotopic ratio $^{13}C/^{12}C$ in order to assess whether biology is controlling the reactions producing methane.

The search for extinct life on Mars by robotic means must be part of a larger effort to thoroughly characterize the histories of water, carbon compounds, and other biogenically important elements on Mars. A properly conducted series of expeditions will enable an understanding of when and how life could have begun on Mars or, if no evidence of life appears, why it never occurred. A series of spacecraft missions beginning in 1996 and culminating in the Mars Exploration Rovers *Spirit* and *Opportunity* (Chapter 15) set the stage for a more ambitious program over the next decades (Chapter 15) to begin to answer these haunting questions.

12.6.5 Earth

Earth is home to an abundant variety of biological organisms, all of which are carbon based and utilize the nucleic-acid-based molecules DNA and RNA to record and transfer the blueprint for structures and metabolic processes. With the exception of viruses and the even-simpler virons, both of which rely on DNA-based cells for their continued existence, all Earth life is organized through the fundamental unit of the cell. Two types of cells exist: prokaryotic (simple) and eukaryotic (complex) cells. Earth life uses a variety of processes to capture energy from the Sun and Earth's interior and to store it in sugars and then high-energy phosphate bonds to power production of proteins and replication of genetic materials. Most types of eukaryotic cells must employ molecular oxygen in metabolic processes because of their high demand for energy; free oxygen, in turn, is produced by photosynthesis in plants and certain bacteria.

At the most fundamental level, Earth life can be divided into three groups: the prokaryotic *bacteria* or *eubacteria* (true bacteria), *archaea* (literally, old ones; also prokaryotic), and *eukarya* (organisms composed of eukaryotic cells, including fungi, plants, and animals). By comparing the resemblances between the RNA in various organisms, one can construct a tree on which the relative distances between organisms indicate their degree of commonality, as shown in Figure 12.10.

The archaea are perhaps the least familiar type of organism to most of us. This domain includes organism such as the *methanogens*, which produce methane in an anaerobic energy cycle, the *halobacteria*, which survive in extremely salty environments, and the *sulfothermophiles*, which live in high-temperature environments such as submarine vent areas and rely on sulfides for their energy source. There is a deep evolutionary gap between these cells and the true bacteria, and it has been popular to argue (hence the name) that the archaea are the most primitive organisms on Earth. However, the French biologist P. Forterre and others have asserted that the archaea show evidence of being later adaptations to unique environments rather than progenitors of other organisms. In particular, the transcription of RNA in archaea is done in a fashion more similar to that in eukaryotes than in bacteria, and eukaryotes are almost certainly cells evolved at late times, because of their reliance on oxygen (with a few exceptions). To trace back the origin of life through the phylogentic tree of Figure 12.10 may not be possible because the "universal ancestor" of us all may not be with us today, even among the simplest life-forms.

Ironically, Earth might be the best place to search for alien forms of life, merely because it is relatively accessible. The possibility that an alternative "shadow biosphere" might coexist with life on this planet has been raised by ASU astrobiologist

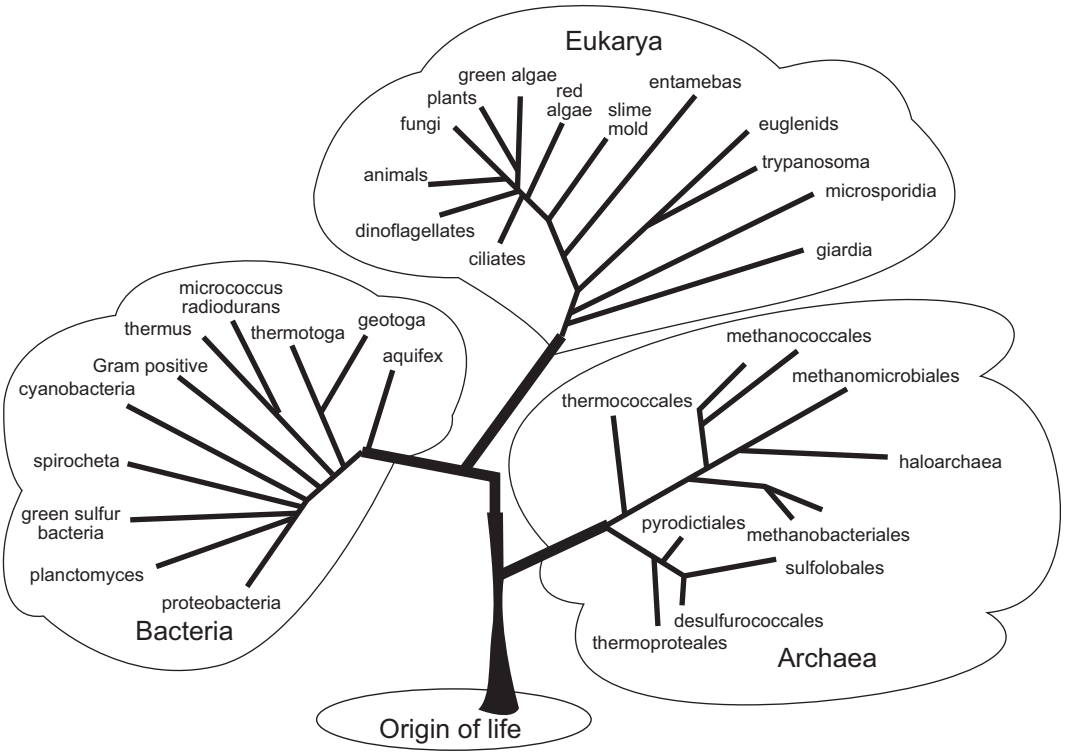


Figure 12.10 The *phylogenetic tree*, or the organization of the major kingdoms of life. The kingdoms are shown grouped within the three domains of life: bacteria, archaea, and eukarya. The tree is constructed by comparing the sequences of nucleotides on RNA molecules in organisms belonging to each of the kingdoms shown. The longer the branch along which a kingdom is located, the greater the difference in genetic sequences from other kingdoms. The differences presumably have resulted from mutations and consequent evolution, which has led to the divergence of life-forms seen today. Hence, the animal kingdom is most closely related to the fungi, and much less so to amoebae with ciliae. The tree has been simplified for clarity by pruning many of the branches. This tree is based on genetic sequences in RNA. From Copley and Summons, 2011.

Paul Davies and colleagues. They point out that potentially habitable environments on the Earth separated from the surface biosphere, such as deep in the crust, might still retain organisms that had a separate origin from our own and therefore should be biochemically different. How biochemically different is a matter of sheer speculation, but the more different such organisms might be from us, the more difficult their detection. Indeed, might there be terrestrial niches so isolated from the surface, and so marginal in terms of habitability, that life there has gone extinct and then begun again countless times over Earth's history? If there are habitable environments and nonhabitable environments in the cosmos, why not "marginally habitable" ones where the transition between chemistry and biology is crossed again and again?

Such questions remain unanswered, because the origin of earthly life is a such daunting question. Nonetheless, we dare tackle it in the next chapter.

Summary

The rock record for the Earth begins about 4 billion years ago, after the Late Heavy Bombardment and after the Earth had acquired an atmosphere and liquid water on its surface. It was in this Archean eon of time that Earth life began, either on our planet directly or transported from a neighboring habitable world (presumably Mars). Evidence of life is definitive in the fossil record some 3.5 billion years ago, but may have appeared hundreds of millions of years earlier. The essential workings of life include small organic molecules that serve as the building blocks for polymers the provide cellular structure, function and the information needed to recreate these molecules over and over again. Proteins are the primary structural material for cells, and also control rates of chemical reactions in cells. Proteins are formed of long chains of amino acids, in a specific order such that the chains fold into complex three dimensional shapes that determine their function. DNA contains the information needed

to build all the proteins used by a particular organism, while a related molecule RNA transfers this information from the DNA to the sites of protein assembly. DNA itself is capable of being copied with a very small probability of error, but the errors that do occur are the raw material for evolution – the development of new species from old – molded, however, by natural selection. Life on Earth is composed of two types of cells: the simpler prokaryotes, and the more complex eukaryotes of which our cells are an example. Eukaryotic cells are capable through their mitochondria to utilize molecular oxygen in their metabolism, with a much higher energy yield than more primitive anaerobic processes of energy generation in cells. Energy for the complex web of life is derived mostly from sunlight through photosynthesis, in which plants and some bacteria produce usable food from water, carbon dioxide and sunlight. A more primitive and less efficient means of generating organic molecules, chemisynthesis, probably predated photosynthesis. The energy produced from metabolic processes (consumption of food) is stored in the phosphate bonds of simple molecules within cells, ATP, and the exchange between ADP and ATP is fundamental to life on Earth. All life on Earth is biochemically the same; drastic changes in the key elements involved in life are difficult to imagine because we have no model for such changes. Water seems particularly suited as the liquid medium for life, but other liquids might work with alternative biochemistries. Beyond our Earth a number of environments might host life: the subsurface of Mars, a liquid water ocean beneath the ice crust of Jupiter's moon Europa, pockets of salty water within Saturn's moon Enceladus, and perhaps even the hydrocarbon lakes and seas on Saturn's giant moon Titan. Claims of evidence for fossil life on Mars are intriguing but not, at this point, compelling. The emission of methane from beneath the surface of Mars might, however, signal either subsurface life or a habitable environment there. On Earth, the biochemical similarity of all known life does not close the door on a speculative possibility that we share our planet with a biota so alien to us that it must have had a separate origin.

Questions

1. Can you think of physical evidence that living processes require a creative spark beyond the chemistry of the nonliving world?
2. Why does DNA use only four distinct nucleic acid bases? Could there be an advantage in a system using six or even eight, such bases?
3. Make a list of the advantages and disadvantages of water as the liquid medium for biochemistry. Using the reference list in this chapter, investigate other possible fluids.
4. Another potential target in the search for life is Saturn's moon Enceladus, which appears from *Cassini* data to have a reservoir of salty liquid water under its surface, and geysers of water (Figure 12.11) emitted from the surface appear to contain organic molecules. Using the reference list below or your own literature search, discuss the case for Enceladus as a potentially habitable world.

Figure 12.11 *Cassini* image of the geysers emitted from Saturn's moon Enceladus. NASA/JPL/SSI image.

General reading

Coustenis, A. and Taylor, F. 2008. *Titan: Exploring an Earth-like World*. World Scientific, Singapore.

Kasting, J. M. 2009. *How to Find a Habitable Planet*. Princeton University Press, Princeton.

Pappalardo, R. T., McKinnon, W. B., and Khurana, K. 2009. *Europa*. University of Arizona Press, Tucson, AZ.

References

Bada, J. L., Glavin, D. P., McDonald, G. D., and Becker, L. 1998. A search for endogenous amino acids in Martian meteorite ALH84001. *Science* **279**, 362–5.

Bell, M. S. 2007. Experimental shock decomposition of siderite and the origin of magnetite in ALH 84001. *Meteoritics and Planetary Science* **42**, 935–49.

Benner, S. A., Ricardo, A., and Carrigan, M. A. 2004. Is there a common chemical model for life in the universe? *Current Opinions in Chemical Biology* **8**, 672–89.

Chyba, C. and Phillips, C. B. 2002. Europa as an abode of life. *Origins of Life and Evolution of the Biosphere* **32**, 47–68.

Copley, S. and Summons, R. 2012. Terran metabolism: the first billion years. In *Frontiers of Astrobiology* (eds. C. Impey, J. Lunine, J. Funès). Cambridge University Press, Cambridge UK. In press.

Davies, P. C. W., Benner, S. A., Cleland, C. E. *et al.* 2009. Signatures of a shadow biosphere. *Astrobiology* **9**, 241–9.

Formisano, V., Atreya, S., Encrenaz, T., Ignatiev, N., and Giuranna, M. 2004. Detection of methane in the atmosphere of Mars. *Science* **306**, 1758–61.

Glein, C. R. and Shock, E. L. 2010. Sodium chloride as a geophysical probe of a subsurface ocean on Enceladus. *Geophysical Research Letters* **37**, L09204.

Jull, A. J. T., Courtney, C., Jeffrey, D. A., and Beck, J. W. 1998. Isotopic evidence for a terrestrial source of organic compounds found in Martian meteorites Allan Hills 84001 and Elephant Moraine 79001. *Science* **279**, 366–9.

Lunine, J.I. 2010. Titan and habitable planets around M-dwarfs. *Faraday Discussions* **147**, 405–418.

Morrison, R. T., and Boyd, R. N. 2008. *Organic Chemistry*, 6th edn. Prentice Hall, New York.

Mumma, M. J. *et al.* 2009. Strong release of methane on Mars in northern summer 2003. *Science* **323**, 1041–5.

Pace, N. R. 1997. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–40.

Parkinson, C. D., Liang, M.-C., Yung, Y. L., and Kirschvink, J. L. 2008. Habitability of Enceladus: Planetary conditions for life. *Origin of Life and Evolution of the Biosphere* **38**, 355–69.

Schopf, J. W. 2006. Fossil evidence of Archean life. *Philosophical Transactions of the Royal Society, London* **B361**, 869–85.

Schwartz, J. H. and Maresca, B. 2006. Do molecular clocks run at all? A critique of molecular systematics. *Biological Theory* **1**, 357371.

Shapiro, R. and Schulze-Makuch, D. 2009. The search for alien life in our solar system: strategies and priorities *Astrobiology* **9**, 1–9.

Tsokolov, S. A. 2009. Why is the definition of life so elusive? Epistemological considerations. *Astrobiology* **9**, 401–11.

Westheimer, F. H. 1987. Why nature chose phosphates. *Science* **235**, 1173–8.

Woese, C. R., Kandler, O., and Wheelis, M. L. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the USA* **87**, 4576–9.

Wolfe-Simon, F., Blum, J. S., Kulp, T. R. *et al*. 2010. A bacterium that can grow by using arsenic instead of phosphorus. *Science* DOI: 10.1126/science.1197258.

Zolotov, M. Y. and Shock, E. 2000. An abiotic origin for hydrocarbons in the Allan Hills 84001 martian meteorite through cooling of magmatic and impact-generated gases. *Meteoritics and Planetary Science* **35**, 629–38.

13

The Archean eon and the origin of life II Mechanisms

Introduction

Having covered in the previous chapter the basics of present-day living organisms and considered the limitations of life in terms of terrestrial and extraterrestrial environments, the present chapter addresses some of the issues surrounding the origin of life. We begin with general considerations about living processes and their relationship to the natural laws that govern the workings of the universe. In particular, self-organization seems to be a property of complicated physical systems, and computer simulations of such systems suggest the kind of bootstrapping necessary to build well-controlled biochemical processes from simpler suites of chemical reactions. We then move to more specific ideas about how life might have begun and examine the issue from two somewhat different points of view: that the origin of life lay in the primitive mimicking of cellular processes (the vesicle model) or that the essential point of origin lay in an RNA or slightly more primitive genetic-coding molecule (the RNA world).

13.1 Thermodynamics and life

Thermodynamics, introduced in Chapter 3, is the study of energy transfer in macroscopic systems. A fundamental principle that governs the transfer of matter and energy in both natural and artificial systems is called the second law of thermodynamics. It is most precisely expressed mathematically but, in words, it says that the capacity for a system to do useful work (move something) decreases with time, unless usable energy is pumped into it. Looked at another way, systems tend to become more disordered with time. An engine takes a hot fluid and converts this to motion of some kind; the waste heat is dumped into a cold reservoir. Over time, the hot reservoir cools, the cool one warms, and the efficiency of the engine to keep the moving parts moving decreases. Only by heating up the hot reservoir (adding some energy from outside) can the engine be maintained at a given efficiency; but then the device heating up the reservoir must itself be stoked or it will run down, too.

A measure of the ability of a system to do work on its environment is called *entropy*. The more entropy, the more disordered the system and the lower the system's capacity to do work. The entropy of a system increases with time unless an external device is applied to pump more usable energy into it. But then, that external device gains entropy. In fact, a fundamental principle of thermodynamics is that the total entropy of the universe is increasing with time, mostly thanks to the nuclear fusion going on in stars, which makes them produce photons, and hence entropy. One situation in which entropy is not increasing, but is constant instead, is that of a system in equilibrium with its surroundings; this is a state in which there is no tendency to exchange matter or energy with the surroundings. Can work be done in such a state? No – hence systems in which entropy is not increasing can do no work. A big tank of water can do no work if it is in contact with a body of water at the same temperature and pressure, that is, it is in equilibrium with its surroundings.

The second law, along with other laws and corollaries, is fundamental to the design of machines in our civilization and is observed repeatedly in the natural world around us. A chemical reaction converts fuel in a camping stove to carbon soot and various gases, releasing heat and driving the system toward a more disordered state; while it does so, food may be cooked. After the fuel is converted to gas and soot, it is hard to use these products to heat after-dinner coffee! Snow taken from the frigid heights of a high mountain down to the surrounding lowlands melts quickly; once melted, it requires a refrigerator (powered by electricity) to make snow or ice again. An egg may be scrambled and cooked to taste using various sources of energy; to reconstitute the original egg from the scramble is a bit more challenging.

What is the physical origin of the second law? It arises from the fact that the macroscopic (large-scale) world that we see actually is composed of a very large number of microscopic

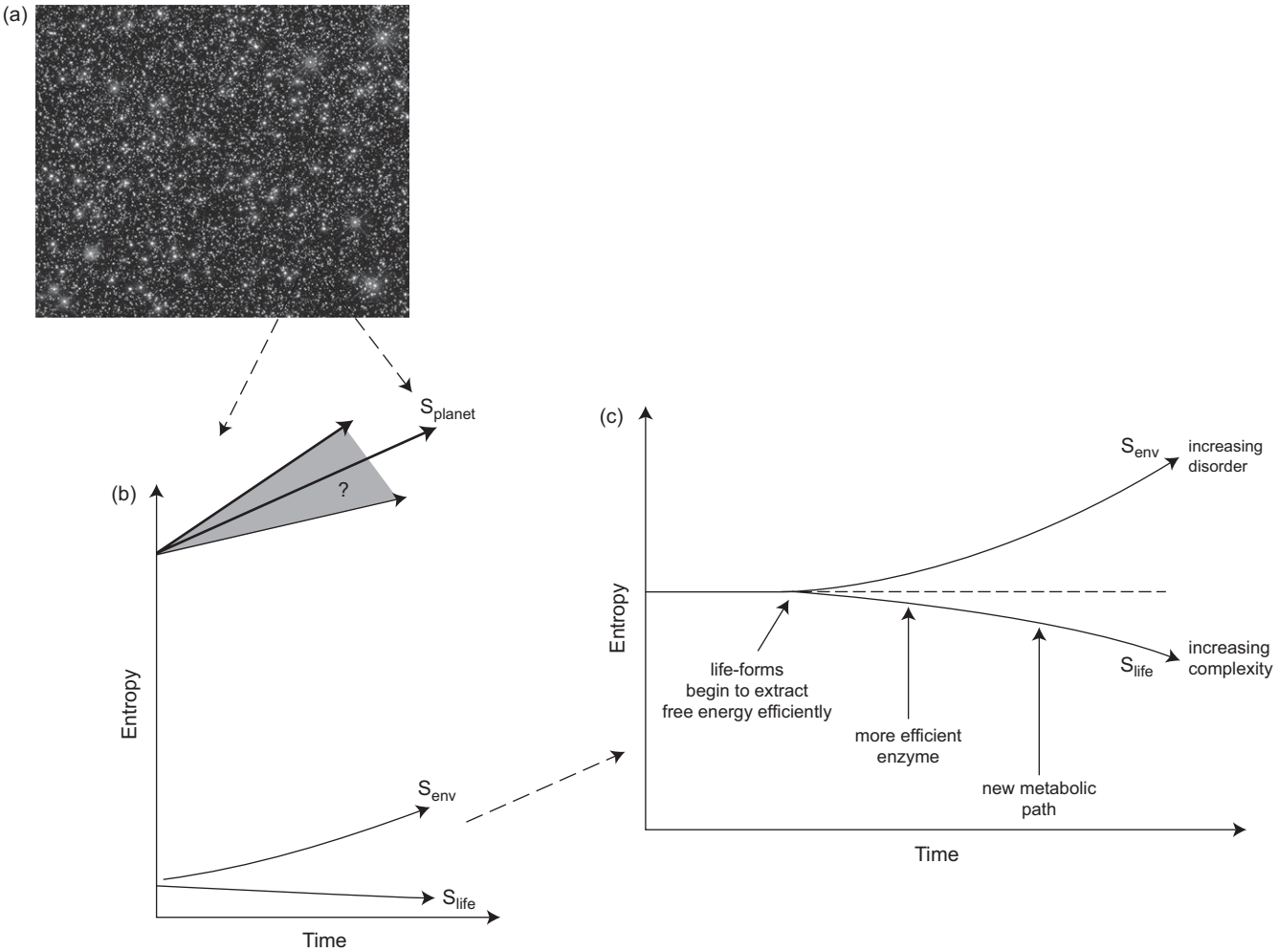


Figure 13.1 Stellar nucleosynthesis is a primary producer of entropy in the modern universe (a), and provides a source of so-called "free energy," which keeps planetary surfaces away from equilibrium. As Earth is warmed by the Sun and radiates energy in the form of heat, it too generates entropy, but life that is hosted on its surface, considered without regard to its environment, would seem to have decreasing entropy with time (b). When considered as a system coupled to its immediate environment ("env"), however, one sees that the net change in entropy of life plus its immediate environment is positive (c). Even life existing in the crust of the Earth, heated by radioactive decay of the elements, uses energy that at its source comes from the fusion of elements in stars, since it is there the elements are made. Figures from C. Lineweaver. For color version see plates section.

particles, and *a given condition of the large-scale world represents an enormous number of possible states in which the microscopic particles may find themselves*. The most probable state of the macroscopic system is the one that expresses the largest number of configurations of the microscopic particles under the given physical conditions – this is the state of equilibrium. Macroscopic systems evolve toward equilibrium as the microscopic particles underlying them wander into more and more accessible states – and hence become less characterizable, or organized, in the process. Equivalently the entropy increases as a macroscopic system evolves toward equilibrium.

It has been said that life is a counterexample to the second law of thermodynamics. Life from seed organizes itself into a more complex state: it grows, becomes stronger, and propagates its own seed. In the fossil record over time, scientists see a remarkable flowering of complex forms from an original rather limited and primitive set of species. Admittedly, all things die, but in the growing and complexification, is life violating the second law? *No*! Life is very definitely providing us with a "living" example of the workings of the second law, and is churning out entropy at a large rate, because it is wholly dependent on its environment for energy, raw materials, and a source for dumping waste (Figure 13.1). Here, it is necessary to elaborate on a corollary of the second law. If a system is brought only a little way away from equilibrium, it will move back to equilibrium in such a way as to produce entropy as slowly as possible. But brought far from equilibrium, the system could operate in very many complicated ways as it moves back toward equilibrium, and the production of entropy speeds up.

This complex behavior as a system is moved far from equilibrium is easily seen in simple everyday life. Open a bottle of soda (or beer) and tilt the bottle just enough to start letting liquid

out. This system is not in equilibrium because the liquid can fall out of the bottle (and do work). It will come out smoothly and predictably. Now open a bottle and turn it upside down; this configuration is much farther from equilibrium. The complex behavior of the liquid as it "glug, glug, glugs" out of the bottle is more interesting and less predictable. And, more entropy is produced.

Systems moved far from equilibrium are at the root of not only complicated behavior, but organized behavior in natural systems. This is a crucial concept! The heating of plains during summertime produces a disorganized stirring of the air around it. The rising air, if it is moist enough, condenses water as it reaches higher, colder altitudes. This condensing process adds more heat to the system and allows the air to rise more rapidly: the system is very far from equilibrium. The eventual result is a beautiful, organized column of cloud, which looks like a giant piece of cotton candy. But even further, the end result of this process – the thunderstorm – produces discharges of electricity that carry enormous power. Surprisingly, a good fraction of the energy of a thunderstorm, originally the energy of the sunlight reaching the ground, is expended in lightning flashes: rather organized electrical energy that is perfectly capable of doing work. Therefore, from the simple, seemingly random process of heating of the ground and transfer of heat to the air, a highly organized system has been created. But entropy is being produced rapidly in this process.

Life is a set of complex physical systems in which chemical and energetic sources are held far from equilibrium. Life produces large amounts of entropy, but because it is so far from equilibrium, organized and complex behavior is to be expected. Living processes do not violate thermodynamics; they simply are maintained far from equilibrium states. This statement can be quantified by comparing living systems with chemical systems that are much closer to equilibrium. The chemistry and energetics of individual living processes are generally well understood today; what is difficult for anyone to comprehend is how these processes interact to produce forms of extraordinary complexity and, in at least one case, sentience.

Life is not in equilibrium with its surroundings; it draws energy from the environment in various forms, and puts waste products back. It is the difference or gradient in a number of properties of the cell relative to the surrounding environment that defines the disequilbrium that simultaneously generates entropy in the environment and the complexity that characterizes life. The understanding of a living organism as a set of interlinked systems held far from equilibrium by virtue of free energy in the environment is essential to approaching the question of how life formed. If life is a natural product of the early evolution of Earth, then the essential characteristics of life must be an outgrowth of early Earth systems that were (i) far from equilibrium and, as a result, (ii) self-organizing and (iii) self-complexifying.

It is easier to hold a small system far from equilibrium than a big one, and so, the isolation of a portion of the environment from the whole likely was involved in the origin of life. This isolated piece of the environment had to have a source of energy, and it had to operate in such a way as to build increasing complexity up to that of the genetic coding molecules RNA and DNA. The pathway to such an entity is not known at present. Little pieces of the puzzle appear to be understood, but they are little indeed.

13.2 The raw materials of life: synthesis and the importance of handedness

That the early Earth was rich in the building blocks of life is now generally accepted. After the chaos of the first few hundred million years, a liquid ocean was stable on the cooling crust of the Hadean Earth. Comets striking Earth would have delivered methane, ammonia, carbon dioxide, carbon monoxide, nitrogen, and other more complex molecules. This atmospheric soup, exposed to ultraviolet light or the electrical discharges of lightning bolts, would have been converted into even more complex organic molecules and nitrogen-bearing species called *nitriles*. In a classic set of experiments in the 1950s, Stanley Miller and Harold Urey showed that, if the atmosphere was rich in methane (CH_4) and ammonia (NH_3) and contained not too much carbon dioxide, spark discharges (simulating lightning) could produce amino acids, the building blocks of proteins, from the gas and water.

The laboratory synthesis of amino acids seemed to promise a quick experimental resolution to the origin-of-life question, but two problems arose in the years subsequent to those experiments. First, the atmosphere in the Miller–Urey flask contained primarily reducing gases, an atmosphere with large amounts of methane as is found today on Saturn's moon Titan. However, models and geochemical data suggest, rather definitively, that the predominant molecule in Earth's early atmosphere was carbon dioxide, as is the case today for Mars and Venus. Although organic molecules such as methane likely were present, as was ammonia, they were probably less abundant than assumed in the Miller–Urey experiment. As the amount of hydrogen-bearing organic molecules relative to carbon dioxide is decreased in the Miller–Urey experiment, the amount of synthesized amino acids plummets. It is possible that enough hydrogen-bearing molecules existed to make some amino acids in the early Earth's environment, but probably not in the extreme quantities manufactured in the original Miller–Urey synthesis.

Subsequent analyses of meteorites revealed that they contain small amounts of amino acids as well, apparently synthesized on parent asteroids or in space. Although the different types of amino acids were much broader than we find in living systems today, the essential lesson is that, on the early Earth, chemical molecules up to the complexity of amino acids were made or supplied readily. We might therefore imagine the early Earth's ocean as a soup of organic molecules, including amino acids, drifting from hot environments to colder environments (for example, from submarine vents to the surface), occasionally being disrupted by impact events that still were frequent, and then reforming in the atmosphere or being supplanted by more material delivered by smaller impactors.

The other important problem has to do with the *chirality* or sense of handedness of the amino acid molecules. Figure 13.2 shows two versions, or *enantiomers*, of the amino acid alanine. Each contains exactly the same number of elements with the same types of chemical bonds, and yet they are the mirror image

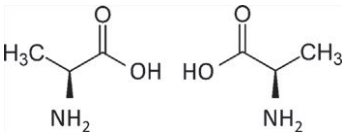


Figure 13.2 Left- and right-handed enantiomers of the amino acid alanine. The left-handed type is referred to as *levorotary* or L-alanine; the right-handed is *dextrorotary*, or D-alanine.

of each other. A molecule that is not superimposable on its mirror image is *chiral*. When a molecule with a definite sense of handedness reacts chemically with one that is symmetric (or otherwise does not have a particular handedness), the left- and right-handed amino acids have similar properties. Likewise, the chemical properties of an interaction between two left-handed molecules or two right-handed molecules are the same. However, neither of these interactions is the same as when a left- and right-handed molecule are interacting with each other. Hence, *the handedness of biological molecules such as amino acids or nucleotides plays a role in their functionality*.

Earthly life has the remarkable property that virtually all proteins are constructed only from left-handed amino acids, whereas the nucleic acids RNA and DNA utilize only right-handed sugars in their structures. Terrestrial organisms cannot utilize right-handed proteins (with a few exceptions) or left-handed sugars in their biochemical processes; they would starve to death if such wrong-handed materials were the sole food source. Yet, abiotically produced amino acids such as those in meteorites and the Miller–Urey experiments are a roughly equal mixture of left-handed (L) and right-handed (D) molecules (there is one meteorite in which there is a modest excess of left-handed amino acids). Furthermore, chemical production of polymers such as proteins or nucleotides does not prefer a particular handedness when the starting molecules are a mixture of L- and D-enantiomers.

In the remainder of this chapter, we use the term chirality to indicate a strong sense of handedness (left or right) in a particular molecular species. *Racemic* means that comparable amounts of left- and right-handed, non-superimposable, molecules with a given chemical formula are present, and *nonchiral* means that the molecule can be superimposed on its mirror image. Chirality is a property; enantiomers are the left- and right-handed versions of a molecule that exhibits chirality.

An important consequence of chirality in biological molecules is that a nonbiological mix of L- and D-type amino acids occurring in meteorites, or in a flask after irradiation (Miller–Urey synthesis), represents more of a problem than a solution to life's origin. How could a particular handedness be selected by prebiological, or primitive-biological, chemistry? We return to this issue later in the chapter because it stands as one of the major challenges to theories of life's origin in which RNA plays an early, primary role.

13.3 Two approaches to life's origin

From here, the road to take is far from certain. The synthesis of amino acids is a far cry from the construction of complex, self-replicating molecules that carry enough information to construct proteins from amino acids. Two approaches to the origin of life from the soup of organics are usually called "metabolism first" and "genetics first" (Figure 13.3). In metabolism first, one argues that certain structures could form spontaneously, capable of isolating parts of the chemical soup from the environment. These *vesicles*, if the right chemicals were present, could have become little factories of increasing chemical complexity, eventually growing and splitting in two but still lacking a reliable (or any) genetic code for reproducing the chemical activity within them. The other approach focuses on the genetic code, in particular RNA, which might have been synthesized in the environment of the chemical soup, and once synthesized, multiplied and co-opted vesicles to form cells. Neither approach yet tells a convincing story, but both have led to some tantalizing suggestions as to how life could have arisen from complex, energetic chemical systems.

13.4 The vesicle approach and autocatalysis

We consider first the vesicles. One of the crucial properties of certain biochemical substances is their ability to enable or speed up reactions, without themselves being expended. This chemical effect is called *catalysis*; some biological catalysts are called enzymes. Catalysis is a common feature in many nonbiological chemical systems, and is essential in biology. A special kind of catalysis is *autocatalysis*, in which a product of a reaction acts as a catalyst in its own production.

Autocatalysis is a process that can lead to complex behavior. Beginning with two chemical substances that tend to react with each other, and supplying enough such *reactants* to maintain vigorous chemical reactions, progressively more complex molecules can be built up in the soup, including molecules that catalyze certain reactions. The key is a continuous supply of reactants, and a source of energy, that is, the system must be maintained far from equilibrium.

Even more significant is that complicated autocatalytic systems, as simulated in computers, have the capacity to increase their level of organization over time. If several sets of autocatalytic cycles are in operation in the same environment, they have the possibility of producing complex chemical species that can couple the sets together and create further organization and complexity. These self-organizing chemical systems increase their network of reaction steps and become more organized so long as energy is available to hold them far from equilibrium.

Scientists have argued that perhaps such autocatalytic sets brought organic chemical systems from the simplest starting components – amino acids and nucleic acid bases, for example – to increasing levels of complexity until primitive proteins and other structures were produced. To do this, the chemical system must have been held far from equilibrium, which means isolating the system from the surrounding environment and enabling energy and reactants to be pumped in and products to be removed. This is where vesicles play a role. In a watery (*aqueous*) medium, certain simple molecules spontaneously form *bilayer membranes* that partition off an "inside" from an "outside." Biological materials such as egg white will do this. Simpler organic molecules called *lipids*, which need not be

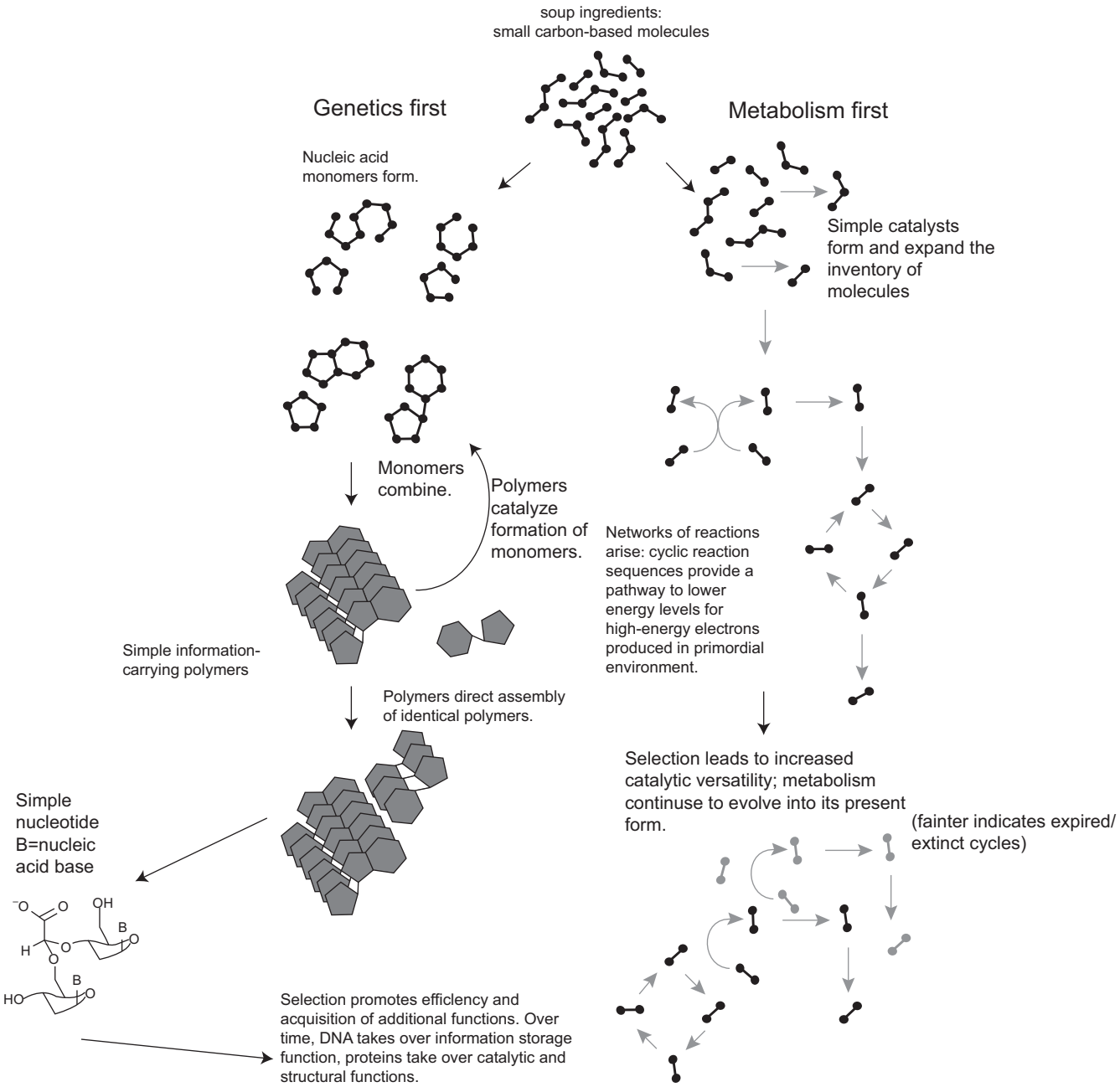


Figure 13.3 Two different conceptual models for the origin of life, "genetics first" (left) and "metabolism first" (right). In each case, one begins with small organic molecules synthesized abiotically on the Earth or even in parent bodies of meteorites. With genetics first, an information carrying molecule arises spontaneously and eventually controls chemical reactions in primitive cells. With metabolism first, undirected networks of chemical reactions – autocatalytic cycles – evolve to progressively higher complexity until they develop information-carrying molecules that eventually control the production of catalysts. Adapted from a figure by Trefil *et al*. (2009), with notional nucleotide from Englehart and Hud (2010).

produced biologically, will do likewise. So, in the early Earth's oceans, it may have been that small environments, microns or less in size, were partitioned off spontaneously by simple organic molecules.

Within the interior of a vesicle, the environment was partially isolated from the outside. It then would be possible to create a small system, out of equilibrium, within which complex chemical reactions, perhaps autocatalytic sets, could have occurred. To do so, we must specify two essential ingredients, namely, energy and a pathway for introducing reactants and removing products.

The spontaneously formed vesicles would not themselves be a source of energy, except for the heat given off by chemical reactions inside them. This heat, though, represents increase of entropy and is not available to do work inside the vesicle. Likewise, heating from the outside is possible but would merely tend

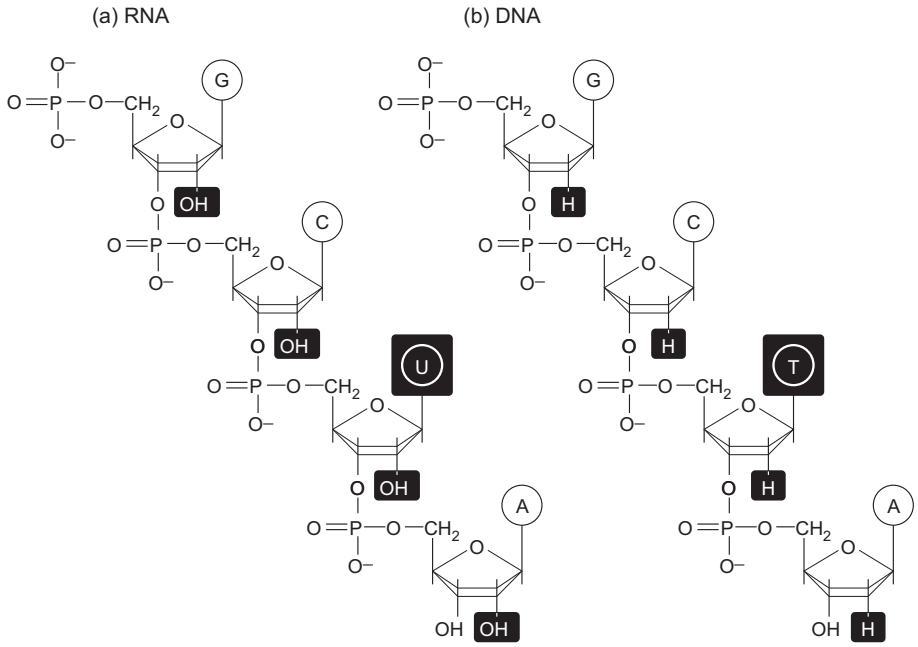


Figure 13.4 (a) Primary structure of the RNA molecule. Nucleotides are labeled within circles by the first letter of their name (for example, G for guanine). Other letters are elements (O for oxygen, P for phosphorus, etc.). Carbon atoms occupy the unlabeled vertices, in accordance with common chemical convention. The straight lines show *single* and *double* bonds, which reflect the number of electrons shared. (b) Structure of a DNA molecule. The differences between RNA and DNA are highlighted on the two structures.

to equalize the inside and outside temperatures – not a promising start for bringing our environment away from equilibrium. One novel suggestion that has been made recently is that certain simple organic molecules, attachable to the vesicle, may have had the capacity for capturing light energy from the Sun and using it to ionize parts of the vesicle membrane. Although speculative, such molecules would *transduce* energy from the Sun and make it available as chemical energy (via the ions formed from the membrane) inside the vesicle.

What about the transfer of reactants and products? This also appears to be possible through a particular set of complex organic molecules that could have acted as channels, filtering some substances through and excluding others. Some preliminary experiments have suggested the possibility that such an attribute might be developed on the vesicles through molecules available in the organic soup of the early ocean.

Although still hypothetical, we have conjectured a vesicle machine that can be charged up, transfer molecules in and out, and serve as an isolated environment for autocatalytic reactions – all of this using molecules plausibly available in the prebiotic environment. The purpose of the machine is simply to make molecules of higher and higher complexity. Whether such molecules eventually would move toward proteins and enzymes is completely unclear. The principle, though, is simple: hold it away from equilibrium and let it get more complex!

13.5 The RNA world: a second option

Although vesicles appear to be a natural and compelling structure for evolving complicated chemical factories, they lack a detailed, formalized set of instructions for producing molecules of the complexity of proteins, and for reliably reproducing themselves. A vesicle that has split off from another, and floated off to a slightly different environment in the early ocean, might well become host to completely different chemical cycles, which fail to sustain autocatalytic sets. Living forms are able to continue their chemical processes from one generation to the next. They also exhibit the ability to self-regulate their internal processes in the face of environmental changes that would completely alter or shut down nonbiological chemical cycles – a capability called *homeostasis*.

13.5.1 The promise: RNA as replicator and catalyst

The genetics-first school holds that the formation of life had as its essential step the formation of RNA from abiotic chemical processes. The RNA would then function as a primitive form of life unto itself, existing perhaps in the early ocean and quickly co-opting protective structures such as lipid-like vesicles. The proponents of this view prefer RNA over DNA because it plays various central roles in all cells today. Not only does it synthesize proteins, but it also primes DNA for replication. In modern cells, DNA is required to produce and regulate RNA, but this may be a later refinement in the evolution of life. In fact, the modification of an RNA molecule to make a single strand of DNA is a relatively minor chemical step (Figure 13.4). There is little argument among biologists that RNA came before DNA.

What made the notion of an RNA world so attractive was the discovery by T. Cech, S. Altman, and colleagues that RNA molecules can act as catalysts, participating in and speeding up the production of nucleic acid sequences and other biological molecules. Although biologically occurring RNA molecules are rather weak catalysts, these abilities can be enhanced and

expanded in the laboratory, by splicing and reproduction of selected sequences on the RNA chain of nucleotides. The resulting modified RNA structures are sufficiently impressive catalysts that it is possible to imagine early RNA as biological catalyst in place of the present-day proteins that function as enzymes. Thus, RNA could have been both the reproductive and the catalytic molecule of the very early stages of life on Earth, acting in autocatalytic cycles to sustain a primitive biology.

13.5.2 The problem: invention of RNA

Most serious is how to put the jigsaw-puzzle RNA together in the first place. To understand whether RNA could have been synthesized in the absence of pre-existing biological molecules, it is convenient to consider the three fundamental chemical parts of an RNA molecule: (i) the nucleic acid bases A, G, C, and U; (ii) a *phosphate group* that contains the element phosphorus and serves as the connector of each of the bases; (iii) the sugar ribose that functions as the binder or backbone of the molecule, so that each nucleic acid base is bound to a sugar and these in turn are attached to each other by the phosphate groups. (As noted in Chapter 12, each unit composed of a ribose, a phosphate group, and a nucleic acid base is called a nucelotide; the polymer composed of a string of nucelotides is an RNA molecule.)

The production of nucleic acid bases by nonbiological means appears to be understood at least in the case of adenine (A), cytosine (C), and uracil (U); there does not seem to be any fundamental hurdle in eventually making guanine (G). Somewhat more difficult is the understanding of how a phosphate group would tend to attach to the right position of a ribose molecule to provide the necessary chemical activity; the same challenge is present in attaching the nucleic acid bases to the ribose. However, one might imagine a random assortment of nucleotide-type molecules, those of which that happened to be configured like an RNA nucleotide possessing a chemical advantage.

The real problem lies in the synthesis and preservation of ribose, with the right chirality. Carbohydrates possessing the formula $C_nH_{2n}O_n$, where n represents a number $1, 2, 3, \ldots$, including the sugar ribose, are readily manufactured by reacting formaldehyde (nCH_2O) with itself. A catalyst is required to initiate the reaction, but this is not a problem. The problem is that ribose is not particularly preferred over other sugars nor is it stable. Hence, an autocatalytic cycle designed to produce large amounts of carbohydrates from formaldehyde will not preferentially make ribose nor preserve it.

One novel suggestion that has been made is that clay minerals may have been involved to concentrate ribose. Clay minerals have ordered surfaces that could form templates, forcing organic molecules that bind to their surfaces to form certain structures. Although no synthesis of RNA has occurred this way, the suggestion is in the right direction: force molecules away from randomly defined patterns to a subset of structures that might allow ribose to form preferentially. This suggestion is also geologically consistent because, with the formation of an ocean in the Hadean era, the environment would have become suitable for formation of clays. One then faces the question of how ribose molecules were maintained against chemical processes that tend to decompose them quickly into a nondescript assemblage of polymeric mixtures. A group led by A. Ricardo at the University of Florida found that the mineral borate tends to stabilize ribose against such decomposition. Hence, particularly on the Earth, where borate-containing minerals should have been common in the crust, repositories of ribose for production of RNA may have been available as raw material for long periods of time.

Yet there are still more difficulties. The ribose produced must have the correct handedness or chirality; on Earth, D-sugars are exclusively involved in living processes. Production of a mixture of D- and L-sugars produces nucelotides that do not fit together properly, producing a very open, weak structure that cannot survive to replicate, catalyze, or synthesize other biological molecules. In fact, the synthesis of the RNA molecule itself is interrupted by mixing nucleotides of different chirality; only in a controlled laboratory experiment or theoretical model can such an assemblage be realized (Figure 13.5).

To create a properly functioning RNA molecule out of a batch of heterochiral L- and D-sugars is a daunting challenge. Two approaches have been pursued. The first is to consider precursor molecules with function similar to RNA but which are much easier to synthesize. The second approach is to understand how prebiological or very primitive biological processes could have selected a particular chirality and allowed its dominance, and hence permitted RNA. In a sense, these two approaches are linked; some precursor chemistry must have operated out of a heterochiral soup prior to the concentration of the D-sugars.

Considering the first approach, it is possible to imagine substituting another sugar for ribose in making RNA, and in particular a sugar that is symmetric and hence nonchiralic. Possible sugars suggested by University of California biologist G. Joyce include glycerol; others have suggested additional candidates such as glucose. Candidates proposed are generally ones that could have been fairly easily synthesized on the early Earth by nonbiological processes, and as sugars, they are capable of binding a nucleic acid base and a phosphate group. However, the properties of the resulting pseudo nucleic acids can be very different. Some have much more flexible structures than RNA, leading to a much greater chance of break up and hence replication or catalysis failure in a fluctuating environment. Others are too stable, and may not catalyze. Finally, many of these substitutes allow not only complimentary pairing (A with U, G with C, as in Chapter 12) but also other pairings (A with C, A with A, G with U, etc.). Under such conditions, the genetic template that sustains a particular kind of chemistry and set of structures is quickly lost after just one generation.

Other possibilities have been conceptualized. For example, amino acids are found readily in meteorites and synthesized under early Earth conditions; could they substitute for sugars as the RNA foundation? Indeed, one can synthesize a backbone composed of glycine (an amino acid) attached to a nitrogen and hydrocarbon unit. This then can attach to a nucleic acid base and a phosphate group to form a nucleotide. The resulting structure is a *peptide nucleic acid*, or PNA. PNAs have been synthesized and shown both to be sturdy and to produce pairing of complementary nucleic acid bases (i.e., A with U, G with C). They could therefore serve as a replicator molecule. However, three open questions remain. Can PNA function as a catalyst? Can one actually induce the polymerization of the amino acid with the nucleic acid bases to make PNA in a plausible prebiotic setting? Is PNA subject to the same chiral restrictions that RNA is? (Recall that many amino acids exhibit chirality.)

(a) (b)

Figure 13.5 The disaster of heterochirality (mixing L- and D-sugars in nucleotides): (a) a normal DNA molecule built of nucleotides of a single chirality; (b) a DNA molecule built with *just one* nucleotide of the opposite chirality (i.e., all D-ribose except for one nucleotide with L-ribose). The one-defect DNA is forced into a much looser structure. The strain in the chemical bonds created by trying to force L- and D-nucleotides together causes bond breakages elsewhere in the structure. The result is a much more open, loose DNA molecule, which is very fragile and thus cannot carry out its templating and replicating functions before falling apart. A similar problem faces the synthesis of RNA from a heterochiralic soup of nucleotides. Photographs reproduced from Avetisov *et al.* (1991) by permission of American Institute of Physics.

The second approach is to understand how the dominance of D-sugars and L-amino acids took place on the early Earth from an initially heterochiralic soup. One would first look to some innate preference for one or the other handedness in the environment or the nature of the molecules. The environment itself yields small effects, which tend to select out one or the other sense of handedness, but some means to amplify the selection must be found. Interestingly, at the subatomic physics level, there is a very small preference for right-handed sugars and left-handed amino acids, the current state on Earth. Such a preference is so small, however, that it cannot *by itself* lead to the distillation of L-amino acids and D-sugars from a heterochiralic soup. Recent analysis of the Murchison meteorite indicates a significant overabundance of some L-amino acids relative to D-amino acids, but the origin of this imbalance and its possible connection to prebiotic chemistry have not yet been explored.

If there is an answer to the development of preferred handedness at the dawn of life, it might well lie in the propensity of complex physical systems to self-organize, as discussed at the very beginning of this chapter. Mathematical models of autocatalytic systems in which polymer production takes place in a racemic environment show that, under certain conditions, the system can exhibit a transition in which the symmetric treatment of D- and L-molecules is broken, and the system rapidly evolves toward a single kind of chirality. No simple physical preference is at work here; instead it is the intrinsic *chaotic* nature of a complicated physical system, held far from thermodynamic equilibrium, that leads to such a self-organizing property. Certainly an autocatalytic system, enclosed in a special environment that allows energy flow and reactants (nutrient) in and products (waste) out, is a candidate to exhibit such behavior. It is in the necessity to invoke such a behavior to make chirally sensitive molecules such as RNA that we might find the combination of the vesicle world and the RNA world to be a requirement for the formation of life.

13.6 The essentials of a cell and the unification of the two approaches

What are the essentials of a cell? Operationally, they are:

1. a dynamic membrane that exhibits fluid-like, flexible motion
2. a set of embedded, membrane proteins that capture energy-bearing molecules (*metabolites*) from the environment, and transport them into the cell
3. a set of enzymes that break down the metabolites and use the breakdown products to construct more membrane, more enzymes, and more genetic material (RNA/ DNA)
4. a genetic string, RNA coded by DNA, which encodes for the set of enzymes
5. a genetic program, DNA primed by RNA, consisting of the set of triggering relations between the various genes. The program will cause the cell to grow, duplicate the genetic-string DNA, and eventually divide when it has gotten large enough, resulting in two cells that will continue to metabolize, grow, and divide.

The chemical vesicle factories embody in a primitive way properties (1) through (3); the RNA world covers (4) and (5). Neither model yields all five properties. It may be that if the origin of life occurred as a natural chemical process on Earth, the first step was the formation of the autocatalytic vesicles, which were short-lived and formed over and over again in different varieties over millions of years – chemical experiments that failed repeatedly. However, at some point a vesicle system exhibited the property of producing polymers of a dominantly single chirality, either sugars or amino acids, and within this system the production of an RNA, a PNA, or other nucleic acid structure (ONA) was enabled. ONA varieties with catalytic capability became coupled into the autocatalytic networks of

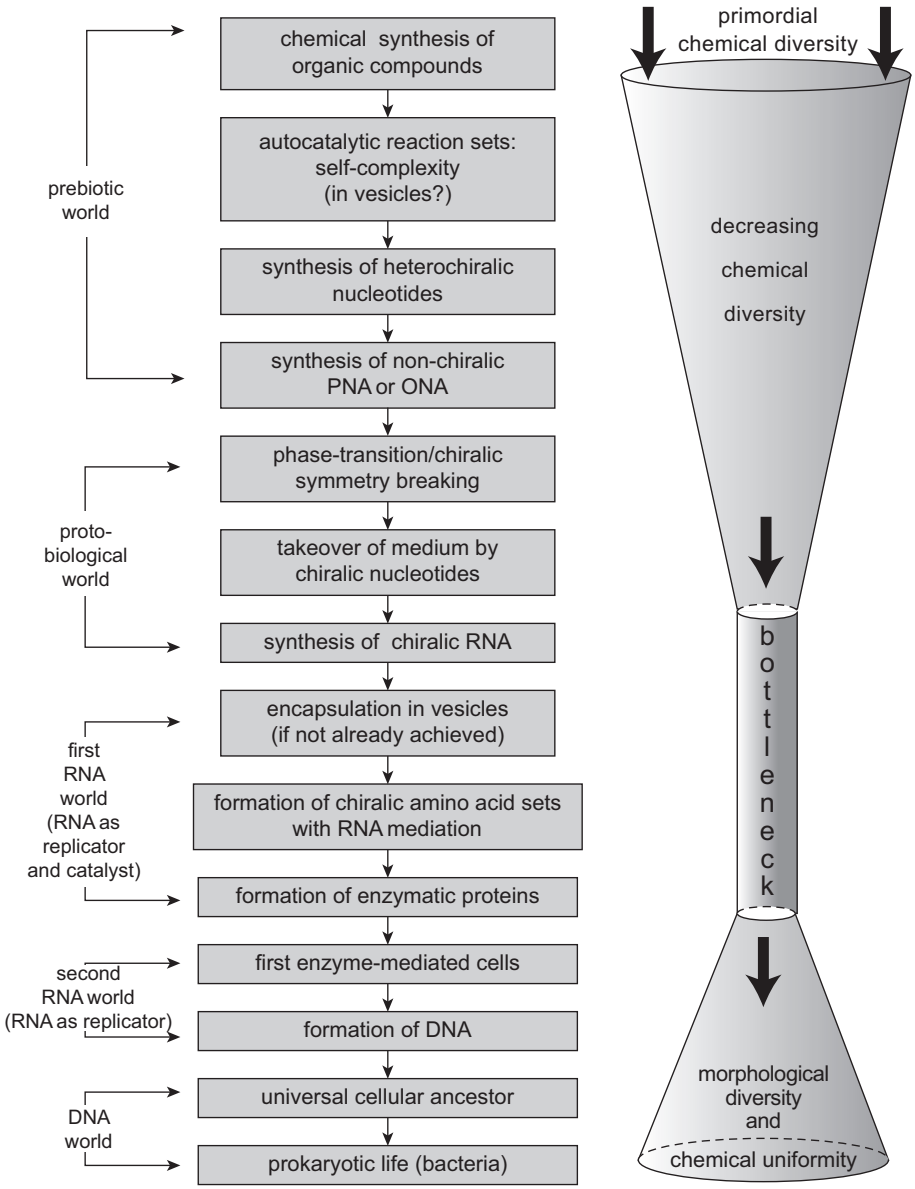


Figure 13.6 One possible schedule of the steps by which life formed. In the left-hand sequence, RNA appears before encapsulation in vesicles, although as the text argues, the reverse might well have been the case. The bottleneck in the origin of a chemically uniform, morphologically diverse biology from a chemically diverse terrestrial environment is illustrated on the right, aligned with the chemical/biological steps on the left. Adapted from Cloud (1988).

some vesicles, and a subset of these used the energy and catalytic properties of the sets to reproduce. Such symbiosis, which is a theme in the evolution of life, could have represented the very primitive precursor to a biological cell. All of the ingredients, (1)–(5), of a cellular structure capable of maintaining and reproducing itself are present in such an RNA-primed vesicle.

Although from here, the formation of DNA is not well understood either, the jump in complexity from RNA to DNA is not considered by biochemists to be as much of a hurdle. Along the way, in some RNA-driven vesicle, DNA may have arisen, and the universal cellular ancestor of all Earthly life was born.

Figure 13.6 suggests two ways to look at the origin-of-life issue. One, on the left, is to try to list the steps, in order, by which life began; this approach is fraught with dissent because we still do not know whether vesicles, RNA, ONA, chirality, or other precursors came first. (For example, Nobel Laureate C. de Duve notes that the development of energy storage in phosphorus-bearing molecules such as *adenosine triphosphate* [ATP] is yet another problem that requires the identification of simpler precursor molecules.) The other approach is to recognize that biology represents a self-controlled selection of a subset of possible molecules out of an enormous range of possibilities. DNA, for example, has four kinds of letters and about 1,000,000 base pairs (ladder rungs) per molecule. The number of possible varieties of DNA molecules then is $4^{1,000,000}$, or 4 followed by one million zeros. *It is the role of enzymes, biological catalysts, to suppress the random nature of chemical reactions so as to preserve and ensure a particular suite of biological molecules*

at the rate needed to sustain the production and replication of the whole system.

In this regard, it is perhaps most instructive to view the right-hand sketch in Figure 13.6. The prebiotic Earth was a system of high chemical diversity, but with an environment that tended to select certain elements and naturally occurring molecules as preferred in increasingly complicated autocatalytic chemical systems. Reaction sets that straddle the barrier between biology and chemistry were still chemically diverse, and likely limited in size and capability to interact with the environment: morphologically (appearance-wise) simple. It is the bottleneck of morphologically simple, protobiotic chemical systems that lies at the crux of understanding how life began. As one kind of reliable protobiochemistry took hold, the chemical diversity of protobiology plummeted, but the success of the system was such as to allow the blossoming of a great diversity of morphologies that functionally allowed different kinds of interactions with a changing environment. Today, biological processes have co-opted most of the available carbon and oxygen in the atmosphere, ocean, and continental surface of Earth, so that the chemistry of these elements is largely limited to the rather uniform and specific biochemical processes that sustain life. Most or all of the other planetary environments in our solar system may never have crossed this bottleneck, but how close they came is an intriguing question (Figure 13.7).

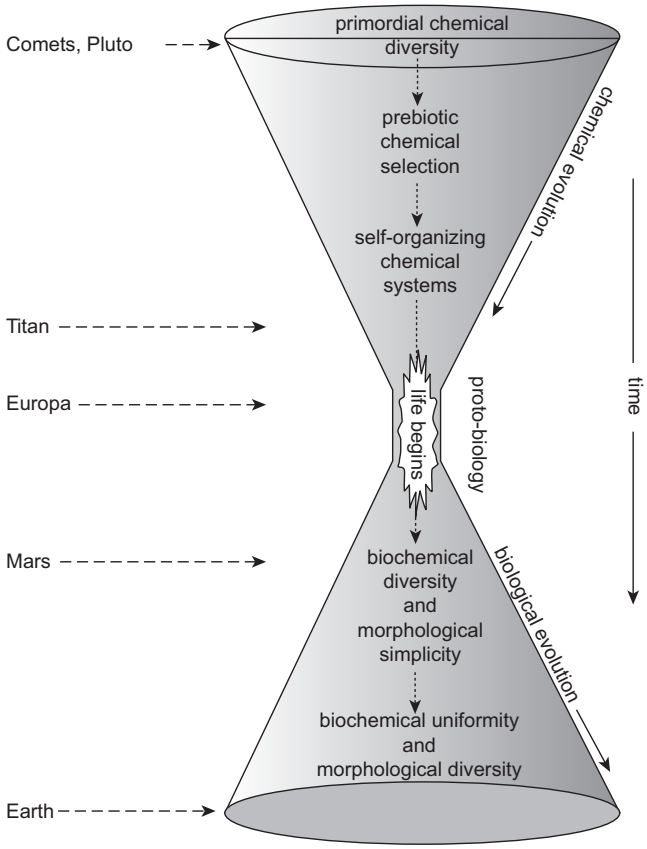


Figure 13.7 Another look at the transition from chemical to biochemical evolution. Conservative guesses as to where various planetary bodies lie on the hourglass are indicated. The cold distant bodies of the outer solar system – Kuiper Belt comets and Pluto – store organic molecules relatively unaltered from interstellar processes. Europa or Enceladus, or both, may harbor life in a subsurface water ocean. Titan's hydrocarbon seas might be sterile or play host to a kind of biochemistry very different from that on Earth. Mars may have had conditions early in its history, and in brief episodes thereafter, capable of sustaining a primitive biota; life might eke out an existence in the planet's water-charged silicate crust. Only Earth, in our solar system, has an atmosphere, ocean, and crust that play host to an extensive biochemistry expressed in the great diversity of life-forms we see today.

13.7 The Archean situation

True life in the Archean, that which left stromatolites and other faint records, consisted of the most primitive type of cells called prokaryotes. The more complex eukaryotes were apparently not present in the Archean, and we tackle their origin in Chapter 19 in the context of the formation of an oxygen-rich atmosphere during the Proterozoic eon. At least in terms of structural complexity of the container of living processes, it does not seem like such a long step from the vesicles of the theorists to the bacterial prokaryotes of the Archean.

When did the prokaryotes form? Certainly before 3.5 billion years ago, based on fossil evidence of photosynthesizing bacteria in Australia, more than perhaps 3.9 billion years ago from isotopic evidence, but after the formation of the liquid water ocean, 4 billion years ago or earlier. The limiting factor may have been the rate of large impacts: if the early Earth environment was rendered unstable by a high frequency of such impacts, there would be no chance for robust vesicles (or RNA) to form. Until the timescale for forming self-sustaining vesicles is understood, we can make no reliable judgment as to when the environment became sufficiently stable. Sometime after the formation of the water ocean, and perhaps sooner rather than later, life appeared on Earth.

What drove organic chemistry toward the threshold to life? Harold Morowitz and colleagues at George Mason University have likened the inevitability of such a bottleneck to other disequilibrium phenomena such as flowing water runs down a hill. While it may flow uniformly as a sheet initially, soon it carves channels, which become gullies and eventually dendritic valleys. Indeed, the energy imbalance represented by the presence of water at the top of a hill is corrected by a process that leads to a kind of spatial ordering of channels separated by ridges. The build up of electric charge differences between a cloud and the ground is not relieved by a uniform flow of electrons; it occurs within a narrow channel maintained far from equilibrium with the surrounding air by the driven flow of the charged particles themselves. One can mentally reduce organic chemical systems to their essence: the charging of electrons within molecules of those systems. The flow of electrons toward lower energy states is not a simple, disordered process if the system is held far from equilibrium by available sources of energy: instead, preferred paths such as particular metabolic cycles may develop. Morowitz and colleagues offer the citric acid cycle, run in reverse so that CO_2 and water are converted to organic molecules with the application of available energy, as a possible example of an early, fundamental metabolic cycle. Whether they are right about the particular metabolic cycles that were foundational,

the general point – that life is an inevitable response of an organic chemical system held far from equilibrium – would argue that life is a common feature of the cosmos and that even in exotic environments such as the hydrocarbon seas of Titan, the complexity and specificity of life should arise from abiotic chemistry.

Summary

The second law of thermodynamics states that entropy, a quantity that measures (inversely) the ability of a system to do work, increases with time in any real-world process. Life is a consequence of, not an exception to, the second law of thermodynamics. Given the large amount of energy available (directly or indirectly) from the Sun to do work, and the presence of nutrients in the environment, living systems represent a high degree of organization but at the same time generate entropy, when the surrounding environment is considered. Therefore, the origin of life need not be seen as a miraculous or even singular event, but rather as the outcome of the natural evolution of organic chemistry in an early planetary environment suffused with energy, organic molecules, and water. Whether this environment was on Earth itself, or elsewhere such as Mars, is not known. Meteorites and cometary debris raining down on the Earth early in its history would have provided our planet with the raw materials for life, and possibly even primitive life itself. The specific steps by which life arose from organic chemistry are not known. Examination of biochemistry reveals that it differs from abiotic organic chemistry in its selective nature. Out of many hundreds of different types of amino acids, only 22 are used by living processes. Life uses, with rare exceptions, only left-handed amino acids and right handed sugars – "chiral" molecules that are not symmetric when reflected in the sense of a mirror. These and other examples reveal the high degree of order (low entropy) and selectivity which characterize living systems. Prior to the appearance of the first self-replicating molecule – be it DNA, RNA or a simpler precursor – organic chemistry could have become ordered through networks of reactions generating their own catalysts – according to the laws of physics and chemistry, but in the absence of "natural selection" that is the hallmark of evolution as discussed in Chapter 18. Once a molecule capable of carrying the information needed to synthesize catalysts appeared, the success of subsequent generations of chemical systems depended on adaptation to the environment. At what stage self-replicating, information-carrying molecules appeared is not known, and two different models for the origin of life have been promulgated: one in which metabolic cycles developed before molecules such as RNA and DNA, and the other in which the sequence was reversed. Regardless, the first self-replicating information-carrying molecule was unlikely to be DNA; RNA was almost certainly its precursor. But RNA is sufficiently sophisticated that it may have had precursors whose presence was not recorded in life as we know it. Vesicles – self-folding organic structures allowing chemical networks to be partially isolated from the environment, may have played an important role, as did perhaps surface chemistry on mineral templates like clays. However it began, life had a toehold on the Earth sometime in the Archean, certainly before 3.5 billion years ago and possibly as early as 3.9 billion years before present.

Questions

1. Imagine a planet with two well-developed biota, one able to synthesize left-handed sugars and use right-handed amino acids, the other synthesizing right-handed sugars and using left-handed amino acids. What kinds of competition might ensue in such a situation? Is it intrinsically unstable, i.e., will one form of life win out?
2. If indeed RNA was the initial genetic encoder and DNA developed later, do you think that some viruses might be remnants of that earlier epoch? Why or why not?
3. Entropy is defined mathematically as the logarithm of the number of possible states accessible to a system, multiplied by a constant (a fixed number). What is the ratio of entropies of an abiotic system that indiscriminantly incorporates sugars of either handedness versus life that uses only right-handed sugars?
4. The sequence of steps toward the origin of life in Figure 13.6 is notional. Construct an alternative, based on the discussion in the text or articles you find in the literature, and contrast it with that in the figure.

General reading

Avetisov, V. A., Goldanskii, V. I., and Kuz'min, V. V. 1991. Handedness, origin of life and evolution. *Physics Today* **44** (7), pp. 33–41.

Bagley, R. J. and Farmer, J. D. 1992. Spontaneous emergence of a metabolism. In *Artificial Life II* (Langdon, C. G., Taylor, C., Farmer, J. D., and Rasmussen, S. eds). Addison-Wesley, Redwood City, California, pp. 93–140.

Levy, S. 1992. *Artificial Life: A Report from the Frontier Where Computers Meet Biology*. Vintage Books, New York.

Morrison, R. T., and Boyd, R. N. 2008. *Organic Chemistry*, 6th edn. Prentice Hall, New York.

References

Ehrenfreund, P. and Cami, J. 2010. Cosmic carbon chemistry: from the interstellar medium to the early Earth. *Cold Spring Harbor Perspectives in Biology* **2**, a002097.

Engel, M. H. and Macko, S. A. 1997. Isotopic evidence for extraterrestrial non-racemic amino acids in the Murchison meteorite. *Nature* **389**, 265–7.

Engelhart, A. E. and Hud, N. V. 2010. Primitive genetic polymers. *Cold Spring Harbor Perspectives in Biology* **2**, a002196.

Lauterbur, P. C. 2008. The spontaneous development of biology from chemistry. *Astrobiology* **8**, 3–8.

Lazcano, A. 2010. Historical development of origins research. *Cold Spring Harbor Perspectives in Biology* **2**, a002089.

Morowitz, H. and Smith, E. 2007. Energy flow and the organization of life. *Complexity* **13**, 51–9.

Orgel, L. 2000. A simpler nucleic acid. *Science* **290**, 1306–7.

Pace, N. 1996. New perspectives on the natural microbial world: molecular microbial ecology. *ASM News* **62**, 463–70.

Schwartz, A. W. 1995. The RNA world and its origins. *Planetary and Space Science* **43**, 161–5.

Ricardo, A., Carrigan, M. A., Olcutt, A. N., and Benner, S. A. 2004. Borate minerals stabilize ribose. *Science* **303**, 196.

Trefil, J. Morowitz, H. J., and Smith, E. 2009. The origin of life. *American Scientist* **97**, 206–13.

Zhu, T. F. and Szostak, J. W. 2009. Coupled growth and division of model protocell membranes. *J. American Chem. Soc.* **131**, 5,705–13.

14

The first greenhouse crisis: the faint young Sun

Introduction

If there is one thing we depend on, it is the assurance that the Sun will shine day after day, year after year, constantly and dependably. We base this sense of certainty on the collective human experience of a constant Sun, and indeed, the concern or even terror that total solar eclipses brought on was a strong motivation for building eclipse predictors such as, possible, Stonehenge (Chapter 2). And yet there is strong evidence from the record of climate, from observing other stars, and from the physics of nuclear fusion, that the Sun has not really shined with constant output over time. Indeed, when the Sun was young, it almost certainly had a lower output than today by a significant amount, leading to what is called the "faint early Sun" or "faint young Sun" problem. This chapter explores the physics of that variation and the implications for Earth's ancient climate.

14.1 The case for an equable climate in the Archean

There is ample evidence that the Archean Earth possessed liquid water. The existence of metamorphosed sedimentary rocks from this period, as discussed in Chapter 11, require erosion by liquid water and deposition in a lake or marine environment. The presence of life itself, recorded through isotopic signatures and fossil evidence, also implies liquid water. As discussed in Chapter 12, we know of no living thing today that can get by without water. Many don't require oxygen (and are poisoned by it), but all require liquid water.

Figure 14.1 summarizes constraints arguing for Earth's mean temperatures being above the melting point of water during the Archean. In Chapter 15, we explore the case for a Martian climate, at the time of Earth's Archean eon, which was warmer than at present (either continuously or episodically). In total, the evidence on Earth and Mars points to planetary climates at least as warm as those experienced today. Surprisingly, as we now show, such climates impose rather strong constraints on the nature of the Archean atmospheres of the Earth and Mars – provided our understanding of the evolution of the Sun is correct.

14.2 The faint young Sun

Simple reasoning about the physics of hydrogen fusion indicates that the Sun was cooler in the past than it is at present. As the Sun converts hydrogen to helium, the mean atomic weight of the atoms in the core goes up (helium is four times heavier than hydrogen), whereas the number of atomic nuclei goes down (four hydrogen nuclei having combined to make one helium nucleus). As the core evolves toward a state of heavier but fewer atomic nuclei, it compresses, forcing the density up. The compression of the core toward higher density, in turn, increases the average temperature of the material as the energy of compression is converted into the random energy of collisions of nuclei.

Finally, the rate of fusion is very sensitive to the temperature, such that small increases in temperature lead to a large increase in the rate of fusion. This is the case because fusion requires a threshold collisional speed in order to allow protons (hydrogen nuclei) to overcome their intrinsic repulsion and bind together (see Chapter 3). Hence, a small increase in the mean speed of collisions (small temperature increase) leads to a very much larger percentage of collisions in which the hydrogen nuclei can fuse to form helium. Therefore, over time, the Sun has gotten more luminous. Computer models predict that, at the time of the early Archean, 3.8 billion years ago, the Sun's luminosity was 75% of the present value, that is, it was 25% dimmer than at present. By the end of the Archean eon, 2.5 billion years ago, the Sun's luminosity was 82% of the present-day value (Figure 14.2).

With less sunlight streaming to Earth in the past, the surface would have been colder than at present. The surface temperature of Earth's oceans today, averaged over their surface and over a year, is 288 K. A very rough guide to what the surface

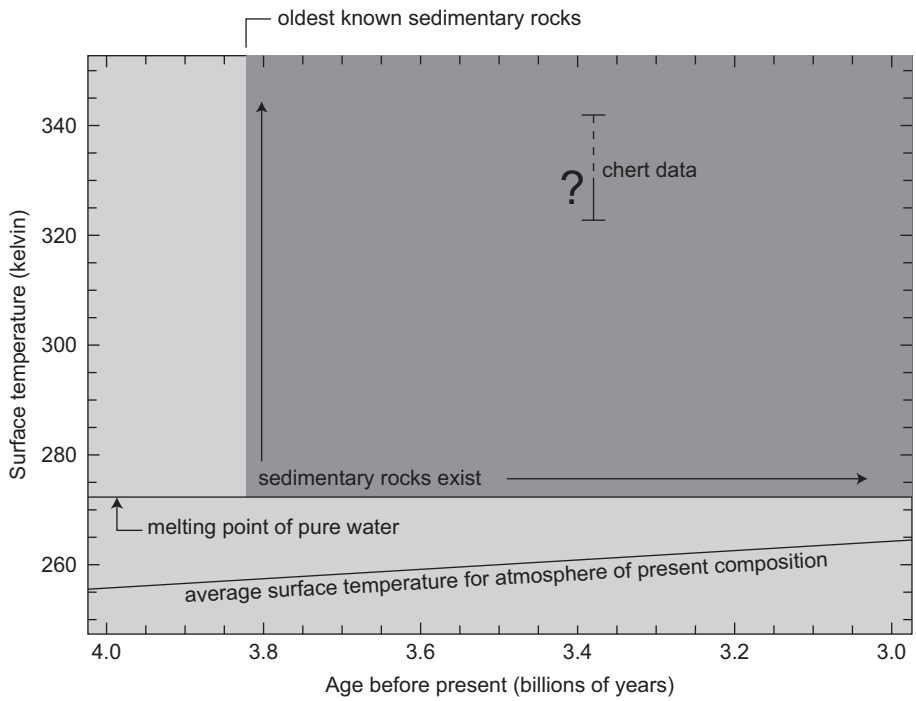


Figure 14.1 Some constraints on Earth's surface temperature during the early to mid-Archean. The line marked "average surface temperature for atmosphere of present composition," derived from Figure 14.2, shows what happens if today's atmosphere is combined with the fainter Archean sun: surface temperatures lie well below freezing. The chert data described in Chapter 6 suggest ocean surface temperatures at 3.4 billion years ago much higher than today's. A more robust but looser constraint is the appearance of metamorphosed sedimentary rocks in the geologic record after 3.85 billion years ago, indicating that widespread liquid water was present and hence that the global mean surface temperature was above the melting point of water.

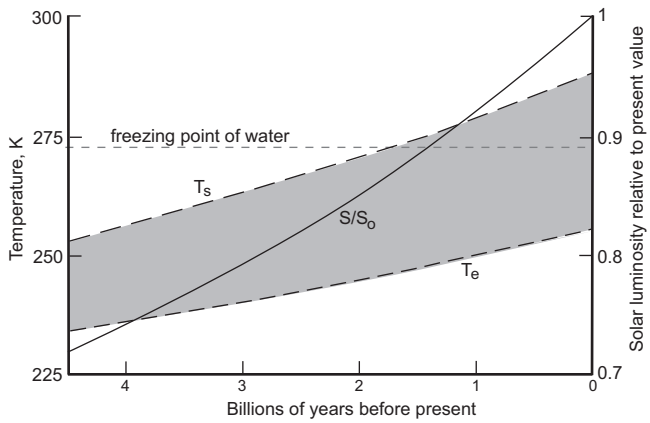


Figure 14.2 The faint young Sun problem. Plotted as a function of time before present are Earth's surface temperature (T_s), its effective temperature (T_e), and the luminosity of the Sun relative to its present value (solid curve). The temperature values are to be read on the left-hand axis; the luminosity refers to the right-hand axis. The surface temperature assumes an atmospheric composition through time identical to the present one, just to illustrate the problem. Under this restriction, Earth's mean surface temperature remains below the freezing point of water (dashed horizontal line) for the first 3 billion years of Earth's history. Reproduced from Kasting (1989) by permission of Academic Press, Inc.

temperature would be for lower solar luminosities (all else kept the same) is given by scaling the temperature to the fourth root of the solar luminosity. (Such a scaling derives from the way in which photons are emitted from a surface that is heated at a given rate.) Hence, at 82% of the solar luminosity, the mean surface temperature is $288 \times (0.82)^{1/4} = 274$ K; for 75% of the solar luminosity, the surface temperature becomes 268 K, below the freezing point of water.

The sensitivity, though, is actually greater than this, because as Earth cools, the atmosphere cannot hold as much water vapor, and this dryness leads to an even lower temperature through the *greenhouse effect*, which we describe in the following sections. Work by Pennsylvania State University atmospheric scientist James Kasting indicates an Earth surface temperature of 255 K at the start of the Archean. Such an Earth could not have had a stable, liquid water ocean. What kept the oceans from being frozen? To answer this question, we need to consider how the atmospheric greenhouse effect works.

14.3 The greenhouse effect

It is a sunny midsummer day and you have parked your car, windows closed, in the asphalt parking lot of your favorite shopping center for an hour of shopping. Upon leaving the building, you notice that the outside air temperature is warm but not broiling. Once the car door opens, though, that familiar blast of heat greets you from the hellishly torrid interior. What happened?

The glass of your car windows and windshield allows plenty of sunlight to get in – glass is transparent at optical (or visible) wavelengths, where most of the Sun's energy is emitted. When the visible photons from the Sun hit the seats, dashboard, and other parts of the passenger interior, they are partly absorbed and then re-emitted as infrared photons, lower in energy but more numerous. Some of the solar photons are instead reflected by bright surfaces and exit through the windows but, even for whiter color schemes, much of the sunlight is absorbed.

The automobile's glass is not transparent at infrared wavelengths: infrared photons are absorbed by the glass on their way out, and partly re-emitted back into the car again. In effect, the heat, in the form of infrared photons, has trouble getting out. (If human eyes were sensitive to light in the infrared rather than the optical, automobile glass would not appear transparent to us.) This situation is not a stable one, because there must be energy balance between the inside and the outside air of the car. As more sunlight streams in and infrared photons are hindered in getting out, the temperature inside the car rises. This increases the flow of infrared photons out of the car, as the inside air temperature rises more and more above the outside value.

Eventually, a balance is reached where the temperature difference between the inside and the outside of the car is enough to balance the free flow of visible photons in, and the arrested flow of infrared photons out. The *greenhouse effect*, then, refers to the increase in temperature of the air caused by the greater difficulty that infrared photons have moving outward to cooler regions, relative to the ease of movement of visible photons. Its efficiency depends on the property of the medium through which the visible and infrared photons move. If they are allowed through the material with equal ease, there is no resulting elevation of air temperature.

For a planetary surface like Earth's, the role of the glass is played by the atmosphere. Sunlight streams down through the atmosphere, some of it reflected by clouds, but a good fraction reaching the surface. The visible photons are absorbed by the ground, people, trees, and buildings. The increased vibrations of the molecules in all these things, engendered by the solar photons, cause emission of heat, in the form of infrared photons, upward through the atmosphere. But the atmosphere, despite being nearly transparent to sunlight (except in cloudy regions), impedes the progress of these infrared photons. They are absorbed and re-emitted many times at many altitude levels in their trip back up through the atmosphere. Photons that are absorbed get re-emitted not only directly upward, but in all directions – up, down, and sideways. Thus some infrared photons, about half in each absorption/re-emission event, find themselves moving downward again – against the flow of energy from the warmer ground to the cooler upper air. The result is an impediment to the outward flow of heat energy (Figure 14.3).

As in the car, there must be a balance of incoming solar energy to outgoing thermal energy. To achieve this, the temperature of each layer in the lower atmosphere must go up, to compensate for the infrared photons turned around and headed downward again. The lower the layer in the atmosphere, the more the temperature increases, because the lower layers are denser and absorb photons more effectively than the thin upper layers. So, the absorption of infrared photons causes the gradient of temperature – its change with altitude – to become steeper.

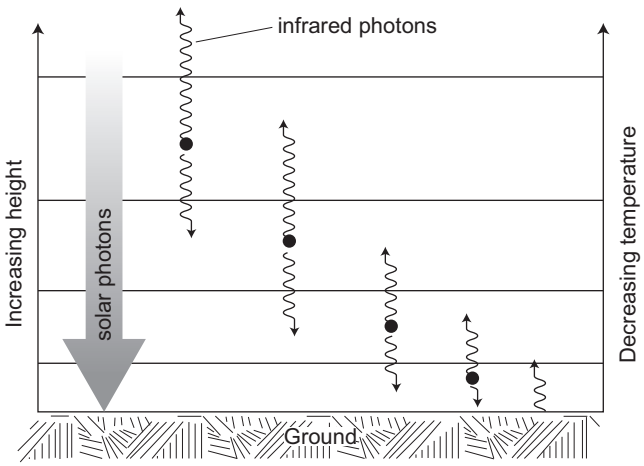


Figure 14.3 Paths of solar photons through the atmosphere (straight arrow) and emitted infrared photons (squiggled lines).

The steeper the temperature gradient, the more efficient the flow of infrared photons outward, until energy balance is achieved. The result is an increase in the temperature at Earth's surface by about 33 K over what we would get if the atmosphere freely let infrared photons escape. For Venus' massive carbon dioxide atmosphere, the temperature *increase* is 500 K!

The actual situation in the atmosphere of Earth is more complicated, as shown in Figure 14.4. Clouds and surface ice very efficiently reflect roughly one-third of the solar photons back outward, so that they never contribute to the warming of Earth's surface and atmosphere. About one-fourth of the solar photons are absorbed directly by clouds or the atmosphere itself. Also, photons usually do not move directly vertically in the atmosphere; there is a spread of photon directions over all angles in the sky, and clouds in particular are very effective at changing the directions of photons (*scattering* of light).

The infrared emission also is somewhat complicated. Because the ground gets several degrees or more warmer than the air immediately above it, turbulent air motions or *convection currents* are set up. The bulk motion, or convection, of the rising warm air moves heat from the ground upward. Convection continues at altitudes well away from the ground: the rising warm air is replaced by cooler air, and the warm air loses heat through radiation of infrared photons in the more transparent higher layers of the atmosphere. In contrast, a hot automobile interior does not cool by convection because the air inside cannot move a significant distance upward; in this respect it is an imperfect analogy for our atmosphere. Therefore, use of the term greenhouse effect for a planetary atmosphere such as Earth's is something of a misnomer, because greenhouses (like a car's interior) cannot cool by convection.

Turbulent motions in our atmosphere are also responsible for forming clouds: as air is raised in the atmosphere, it cools and, as it cools, it becomes less capable of retaining water as a gas. Eventually, the water vapor condenses (much as it does on a cold drink glass) to form clouds. The process of condensation releases some energy in the form of heat, and this must be included in the atmospheric energy balance as well; it is called *latent* heat. Clouds are an important part of Earth's atmospheric greenhouse

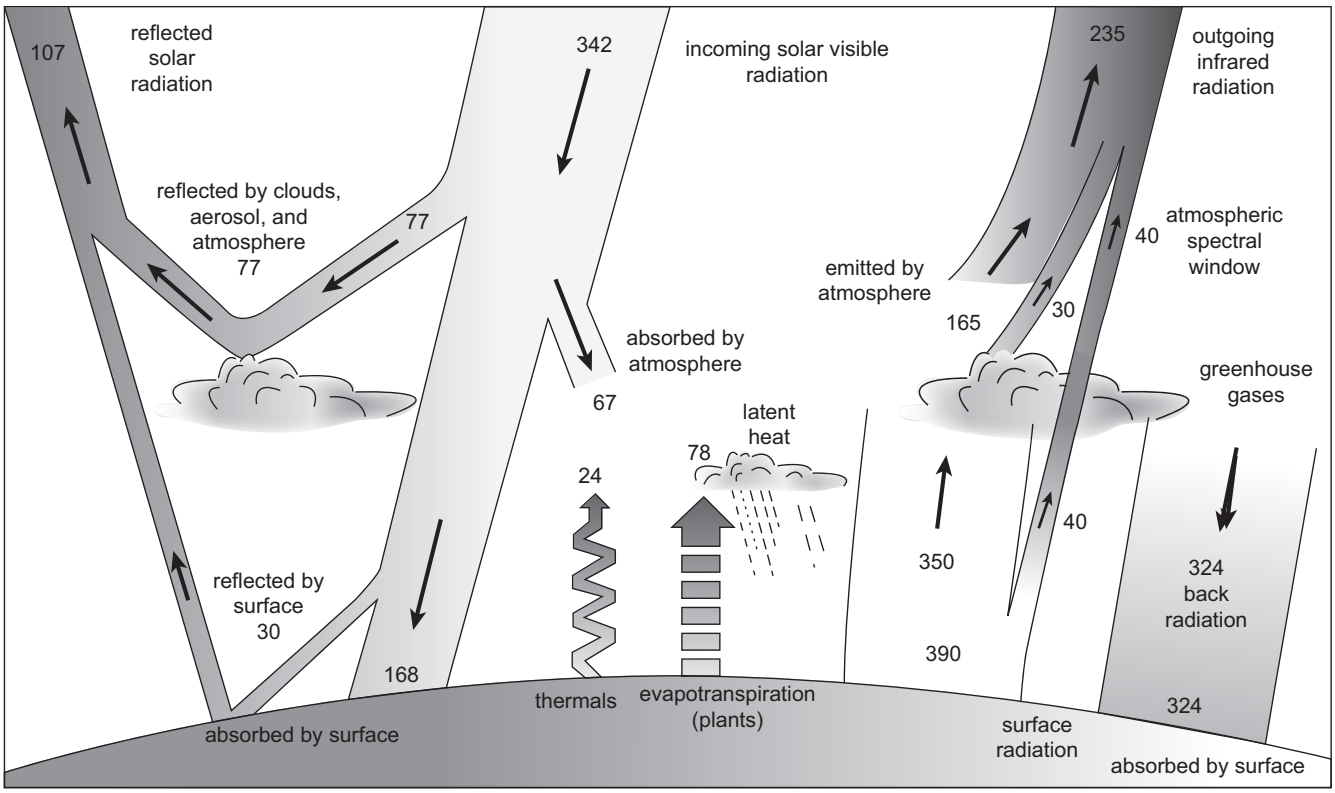


Figure 14.4 Processes affecting the movement of energy in Earth's atmosphere. Solar visible photons are sketched on the left part of the figure, infrared photons on the right, and other kinds of energy transfer in the middle. Numbers show the flux of energy, in watts per square meter, involved in each process; the amount of incoming solar radiation in a square meters per second is an average value for the globe. Latent heat refers to release of heat as water condenses to form clouds; evapotranspiration is the movement of water through plants combined with evaporation back into the atmosphere. Modified from Trenberth *et al.* (1996).

process because they can cool the air (by reflecting sunlight) and heat it (by absorbing infrared photons). The amount of water vapor that the air can hold is a sensitive function of temperature, so that small amounts of warming of Earth can lead to large changes in cloud cover. Also, the rainfall produced by clouds is the principal source of water for land-based life.

Other complications include the daily and seasonal variations in sunlight, the effects of continental landmasses and ocean currents in changing the temperature and moisture content of the atmosphere. These, together with the spinning of the Earth on its axis, generate the very dynamic atmospheric phenomena we call weather. For Earth as a whole, however, the net heating process that warms the surface above freezing is absorption of sunlight at or near the ground, and emission of infrared photons the upward escape of which is impeded by absorption in the atmosphere.

14.4 Primary greenhouse gases

Gases that cause the greenhouse effect must absorb photons in the mid-infrared part of the spectrum, corresponding to wavelengths of light of roughly 10 to 50 microns. The major constituents of the air – oxygen and nitrogen – are poor absorbers. More important are two trace species: water (H_2O) and carbon dioxide (CO_2). Water vapor makes up on average no more than several percent of the molecules in the atmosphere close to Earth's surface, but is highly variable over time and location, depending on the surface temperature and meteorological conditions. Carbon dioxide gas cannot condense to form clouds in the atmosphere, and hence is uniformly distributed around Earth at 0.03%. Methane, nitrous oxide, man-made *chlorofluorocarbons*, and other trace gases contribute as well.

Figure 14.5 shows the wavelengths at which these gases absorb photons. Also given are curves that represent the amount of energy at each wavelength contributed by the Sun, and the distribution of wavelengths at which the energy from the Sun's heating of the ground and air comes out. Note that there is almost no overlap between the wavelengths of solar photons reaching Earth and photons emitted by Earth's warmed surface and atmosphere. (See Chapter 3 for a discussion of the relationship between the temperature of a body and the wavelength distribution of light it emits.) This nice separation of photons into "solar-optical" and "terrestrial-infrared" makes possible the straightforward conceptualization of the greenhouse effect described above.

14.5 Implications for Earth during the faint young Sun era

Returning now to the early Archean, if the amount of sunlight at that time was less than today, assuming (somewhat unrealistically) no change in atmospheric composition, the surface of

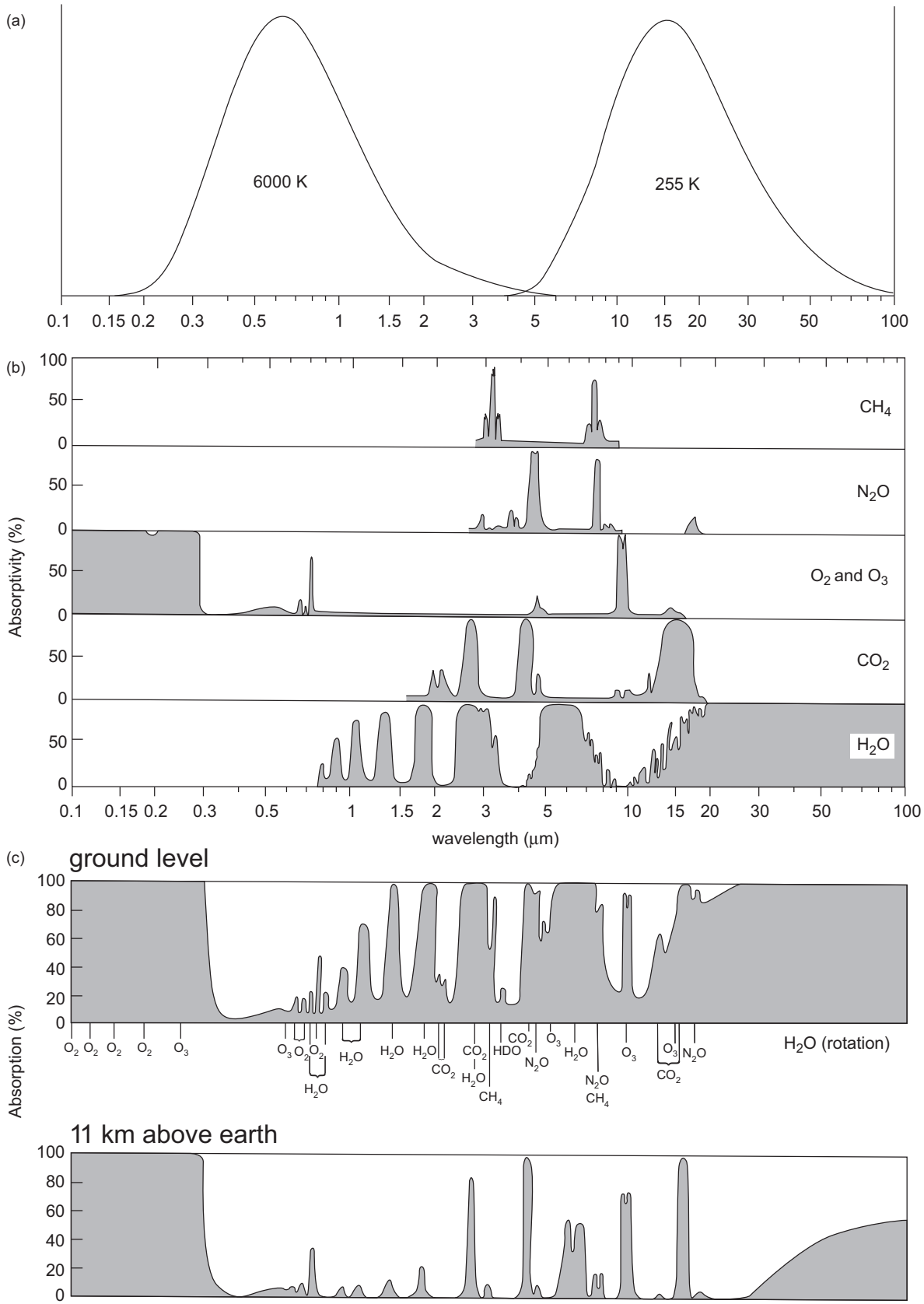


Figure 14.5 (a) Approximate wavelength distribution of photons emitted by the Sun and Earth. (b) Amount of absorption of photons as a function of wavelength, by various gases, through the whole Earth's atmosphere (from the surface upward): 100% absorption means that the atmosphere is completely opaque at the particular wavelength; 0% means the atmosphere is completely transparent. (c) Combined absorption for all gases at sea level, and 11 km above Earth. Notice that, at the surface, the atmosphere is mostly opaque to photons except at optical (0.3–0.7 micron) wavelengths. At a much higher altitude, with less of the gases overhead, the atmosphere is more transparent at infrared wavelengths than it is at Earth's surface. Modified from Peixoto and Oort (1992) by permission of Springer-Verlag.

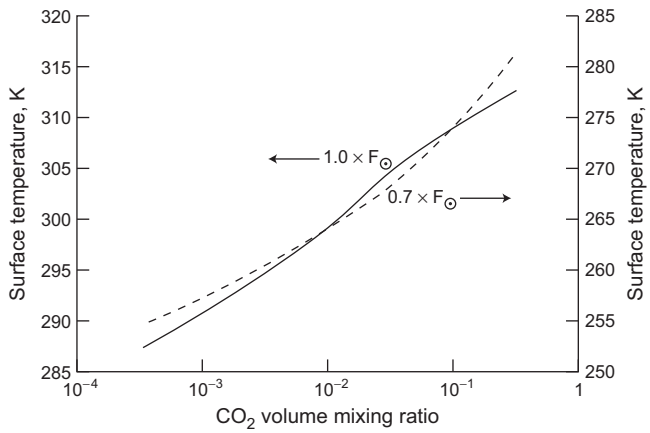


Figure 14.6 Temperature as a function of carbon dioxide abundance in the atmosphere. Solid line is for present-day solar luminosity (scale on left); dashed line assumes a solar luminosity 70% of the present value (scale on right). The air at sea level is assumed to contain one atmosphere pressure of nitrogen, hence, a carbon dioxide mixing ratio of 0.5 means that gas has a pressure of 1.0 atmosphere. Reprinted from Kasting and Ackerman (1986) by permission of the American Association for the Advancement of Science.

Earth would have been below the freezing point of water. But rocks dated at 3.8 billion years ago, when the Sun should have had only 75% of its present luminosity, are metamorphosed from sediments, which strongly suggest that running liquid water was present; other information such as chert data corroborates this. So, the simplest way to obtain an elevated temperature for Earth in the face of a cooler sun is to increase the amount of greenhouse-absorbing gases. Water vapor is controlled by the atmospheric temperature (the higher the temperature, the more evaporation from the ocean), and so, we are not free to simply invoke more water. Carbon dioxide, however, is not so linked and, given that carbon is abundant in continental and oceanic sediments, it is plausible to explore models in which carbon dioxide was more abundant in the atmosphere than at present.

Figure 14.6 is from a calculation by Kasting and Ackerman, using the physics of the greenhouse effect described above, to show the effect of increased carbon dioxide abundance on the surface temperature of Earth. The calculation assumes that there is always 1.0 atmosphere of nitrogen pressure at Earth's surface. The present CO_2 mixing ratio in the atmosphere is 0.0003 in the figure. As more carbon dioxide is added to the atmosphere, surface temperature goes up. The effect is amplified because, as the surface temperature of Earth goes up, more water is evaporated from oceans and goes into the atmosphere, where it further enhances the greenhouse absorption and adds to the temperature increase.

For comparison, two curves are shown, corresponding to the run of temperature with carbon dioxide abundance for the present-day solar luminosity, and for a value 70% of the present, appropriate for the opening of the Archean eon. To obtain an average Earth temperature above the freezing point of water in the Archean requires a carbon dioxide abundance 1,000 times the present value, corresponding to carbon dioxide being one-fourth of the total atmosphere. To obtain a temperature equal to that today, 288 K, requires that carbon dioxide be the dominant gas in the atmosphere, with a pressure well above 1 bar (the current total pressure in Earth's present, mostly nitrogen, atmosphere).

What is the amount of carbon available today that could have been carbon dioxide in the past? An estimate of the carbon buried in sediments on the ocean floor and continents as well as cycling through the upper crust of Earth could yield at least a 60-bar-pressure CO_2 atmosphere, not much less than that of Venus today (see Chapter 15). Hence, there appears to be enough carbon locked in the crust today to maintain above-freezing temperatures in the Archean, if indeed it was in the form of atmospheric carbon dioxide.

Somewhat troubling, however, are the chert data described in Chapter 6. If those data are correctly interpreted, the average ocean temperature in the Archean was higher than today, pushing the required carbon dioxide abundance to even larger, perhaps implausible, values. Because the interpretation of the chert data in terms of surface temperature is highly controversial, what is needed is an independent determination of the carbon dioxide abundance. Although this is not available for the early Archean, it is available for the late Archean. This determination comes from paleosols.

14.6 Paleosols and the carbon dioxide abundance

As rocks are attacked by water during erosion and weathering processes, elements embedded within the crystal structures of the minerals become mobile and may move to other locations. They may form new chemical compounds, and the nature of these compounds can be a sensitive function of the ambient conditions – including the atmospheric composition. *Paleosols*, weathered rocks (that is, soils) that are preserved through burial and hardening (without extensive metamorphism), may preserve an indirect record of the composition of the atmosphere at the time the weathering occurred.

At present, molecular oxygen is the second most abundant gas (21% of the total) in our atmosphere. Its presence and reactivity have a profound effect on rock weathering; in particular, iron readily combines with oxygen to form various iron oxides. Rocks occurring prior to about 2 billion years ago show evidence of ferrous iron (FeO) being mobilized by weathering processes. This relatively oxygen-poor compound of iron requires that the atmospheric oxygen abundance at the time be very low, a fact of profound importance that we consider in Chapter 17. What is important here is that the FeO is dissolved readily in water, mobilized, and becomes available to combine with other materials, the choice of which is sensitive to the atmospheric composition.

In particular, some iron-rich basalts, weathered to soils in oxygen-poor conditions (Archean to early Proterozoic time), lost much of their iron through reaction with water. Some of the iron probably ended up in the local ground-water table, but some reacted with silicates to form iron silicates such as *greenalite* [$Fe_3Si_2O_5(OH)_4$]. The formation of such silicates would not be possible if large amounts of carbon dioxide existed in the atmosphere. The carbon dioxide, diffusing into the soils,

would force the formation of an iron carbonate such as *siderite* ($FeCO_3$). Thus, the absence of iron carbonate in paleosols formed in oxygen-poor conditions sets a limit on the abundance of carbon dioxide.

Analysis of paleosols from Canada, formed 2.2 billion years ago, shows no evidence for iron carbonates; instead iron silicates are present. Thus, an upper limit on the carbon dioxide in the atmosphere at the time of formation should be obtainable. The formation of iron carbonates is sensitive to temperature, and the carbon dioxide concentration in the soils was not exactly the same as that in the atmosphere. Despite these uncertainties, it is possible to use laboratory experiments and theoretical calculations to estimate a carbon dioxide upper limit. A maximum of 3% carbon dioxide in the atmosphere at the time of the paleosol formation, or 100 times the present value. This upper limit is obtained by assuming a soil temperature of 300 K; lower temperatures yield lower upper limits because of the dependence of the chemistry of the iron carbonate formation on temperature. The limit falls to 1 to 10 times the present atmospheric value by about 1 billion years ago.

Note carefully that the number derived is an upper limit to the carbon dioxide concentration – it could have been much lower. Nonetheless, what is important about the result is that even 3% carbon dioxide would have been insufficient to produce a surface temperature of 300 K, given the amount of solar luminosity (80% of the present value) some 2.5 billion years ago. Assuming a lower surface temperature does not help, because that in turn lowers the carbon dioxide upper limit from the paleosol data.

Could it be that enhanced carbon dioxide was not the sole contributor to a thicker greenhouse atmosphere? Other gases, particularly methane, might have been present to contribute additional infrared absorption. Some models of the early Earth atmosphere predict concentrations of methane between 100 and 1,000 times larger than the 1.6 ppm of today. Methane is easily broken apart by sunlight in the presence of oxygen in the modern Earth's atmosphere; the lack of molecular oxygen tends to stabilize methane. Another possible greenhouse gas is ammonia, NH_3, which, however, is extremely unstable to breakdown by sunlight. The late Carl Sagan and Chris Chyba (now at Princeton) suggested that, in the presence of more atmospheric methane, a shield of hydrocarbon aerosols (akin to what is found today on Titan) might have blocked ultraviolet radiation from reaching most of the ammonia, thus extending its lifetime. Recent detailed calculation of the properties of such aerosols by Wolf and Toon at the University of Colorado suggest that they could have been an effective shield. Finally, the role of clouds in the energy balance of the Earth's atmosphere also remains to be properly quantified, but this is extremely difficult to do given that their role in the present climate is very difficult to treat: clouds can warm or cool (Chapter 22).

While methane and even ammonia may have played important roles in enhancing surface temperatures during the Archean eon on Earth, the abundance of carbon and the stability of CO_2 makes enhanced carbon dioxide the favorite of most climatologists as the primary mechanism for maintaining clement surface temperatures over the history of the Earth. Analysis of older paleosols, perhaps early Archean when carbon dioxide must have been well above the threshold to trigger iron carbonate formation, would be an important test. Such analysis must await the discovery of very ancient paleosols. However, regardless of whether there is experimental corroboration of enhanced carbon dioxide, a crucial question that must be answered is: if elevated levels of carbon dioxide existed, where did it go? What was the mechanism for evolving the atmosphere from CO_2-rich to CO_2-poor? It turns out that the answer lies in a cycle that is tied to the fundamental geologic process of plate tectonics, a cycle that continues today and that is partly mediated by life.

14.7 Carbon dioxide cycling and early crustal tectonics

14.7.1 Basic carbon–silicate weathering cycle

How did a thick carbon dioxide atmosphere dwindle away over time? Carbon dioxide is taken out of the atmosphere by weathering of rock, and by plants and bacteria during photosynthesis. Much of the biologically trapped carbon continues to cycle through living organisms at the surface. However, some of the carbon dioxide ends up in shell-forming organisms, which die and drop to the ocean floor, effectively removing the carbon from circulation in the biosphere. The chemistry of the weathering process, which depends critically on rainwater, goes as follows.

The breakdown of silicate rocks by the weathering action of rainwater is very efficient, because CO_2 gas dissolves in the rainwater to make a weak acid that can attack the rock chemically. This yields, among other products, silicon dioxide (SiO_2, the basic molecule of which quartz is made), hydrogen carbonate ions (HCO_3^-, where the negative sign indicates that the ion is negatively charged), and doubly charged ions of calcium (Ca^{2+}). These ions are quite reactive, and are used by shell-forming organisms to make calcium carbonate ($CaCO_3$) shells. The shells, along with silica (opal) shells made by other organisms, are preserved as thick layers of sediment on the floors of lakes, seas, and the ocean. The coral reefs are perhaps the most spectacular example of structures formed by deposition of calcium carbonate.

How long would it take this process to remove all carbon dioxide from the present atmosphere? Calculations show that, at current weathering rates and with the present mass of biota in the oceans, the removal time is less than one million years – only 0.02% of the whole history of Earth. With most of the carbon dioxide gone, the oceans would freeze over very quickly. Something else must happen, both today and in the past, to release the carbon dioxide from the calcium carbonate and return it to the atmosphere.

That "something" is plate tectonics. The ocean floor is continually being recycled back into the mantle at subduction zones, and the carbonate-laden sediments are carried with it. Much of the ocean-floor material subducted at trenches is melted at temperatures well over 1,000 K. Carbon-bearing materials, such as the calcium carbonates, react with silicates at these high temperatures to make calcium silicates ($CaSiO_3$) and CO_2 gas. The gas makes its way back to the surface not at the midocean ridges, but right at the subduction zones where volcanism is occurring. Mount St. Helens, Mount Pinatubo, and other active volcanoes

Table 14.1 Earth's carbon reservoirs (adapted from Falkowski *et al.* 2000)

Reservoir	Amount in gigatons
Atmosphere	720
Ocean	38,000
Carbonates in sediments	≥ 60,000,000
Biomass, alive, in biosphere	600
Dead biomass	1,200
Fossil fuels (oil, goal, gas)	4,100
Kerogens	15,000,000
Inorganic carbon	38,000

belch CO_2 gas from deep within the subducted plates, resupplying the atmosphere.

Carbon is cycled through the atmosphere into the ocean and onto the seafloor, only to be subducted and returned to the atmosphere. This cycle, shown in Figure 14.7, is limited by the time it takes ocean floor (once formed at mid-ocean ridges) to be subducted. At current plate-tectonics spreading rates, the cycle takes about 60 million years; in other words, any given carbon atom in atmospheric carbon dioxide typically will form carbonate, be subducted, and then released again as carbon dioxide gas in a time of order 60 million years.

Table 14.1 lists the known reservoirs of carbon on the Earth today with measured or estimated abundances in one type of unit often used for the carbon cycle: billions of tons, or gigatons. The atmospheric reservoir is mostly carbon dioxide, and the oceanic reservoir is primarily hydrogen carbonates, both of which are dwarfed by the carbonate sediments on the ocean floor. Biomass – carbon in molecules that are part of the biosphere today, either in living or dead organisms, is much less than the carbon in the oceans. Buried carbon exists in several forms. The amount of carbon buried and processed into fossil fuels (Chapter 23) is much smaller than what is believed to have been converted by heat and pressure of deep burial into so-called kerogens – very carbon-rich, hydrogen-poor organic molecules. However, the abundance of kerogens is highly uncertain. Finally, so-called inorganic carbon is that defined to be present in minerals such as limestone, and not containing hydrogen or fluorine. This reservoir is small compared to the carbonate sediments, most of which will become mineralized, with a fraction converted back to carbon dioxide in subduction zones.

14.7.2 Negative feedbacks in the carbon–silicate cycle

Michigan geophysicist J. C. G. Walker proposed that this carbon–silicate weathering cycle might well act as a stabilizing influence on Earth's climate. Because the cycle requires liquid water to dissolve carbon dioxide gas and to effectively weather rock, a fully frozen Earth would have lost the erosive portion of the process. Carbon dioxide gas would not be lost to ocean floor, while more carbon – previously cycled into the crust, or derived from deeper mantle rocks – would continue to accumulate in the atmosphere. This would raise the temperature through the greenhouse effect, until the oceans could melt and liquid precipitation became possible again. (These are two separate conditions; rainfall requires somewhat higher temperatures than does melting the oceans, the freezing point of which is lowered by salts.)

Conversely, higher atmospheric temperatures increase the evaporation from oceans, the amount of cloud, and hence rainfall rates. Higher temperatures also favor rainfall over snow at high latitudes and elevations. These have the net effect of increasing the rate of weathering by rainfall, and hence removal

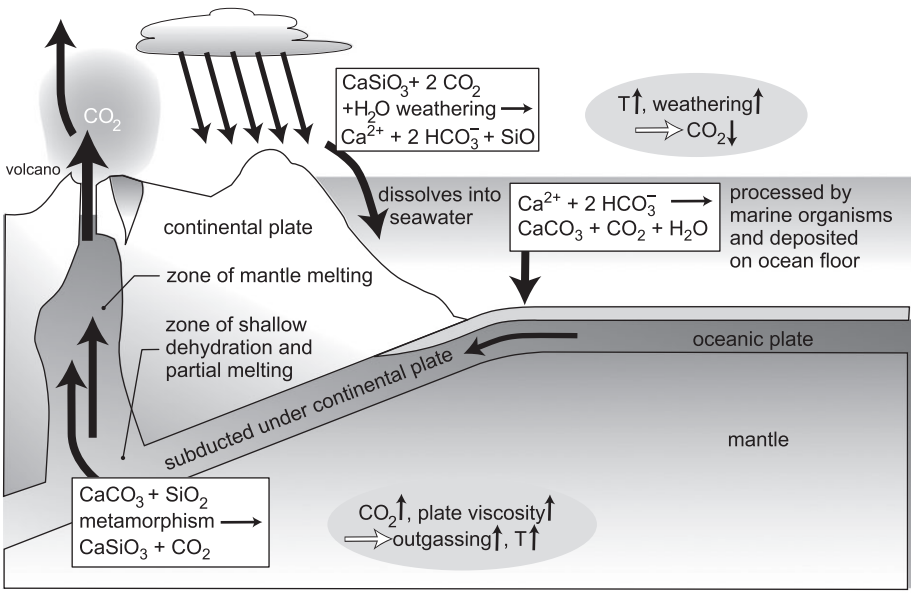


Figure 14.7 Weathering cycle of carbon, silicate, and water.

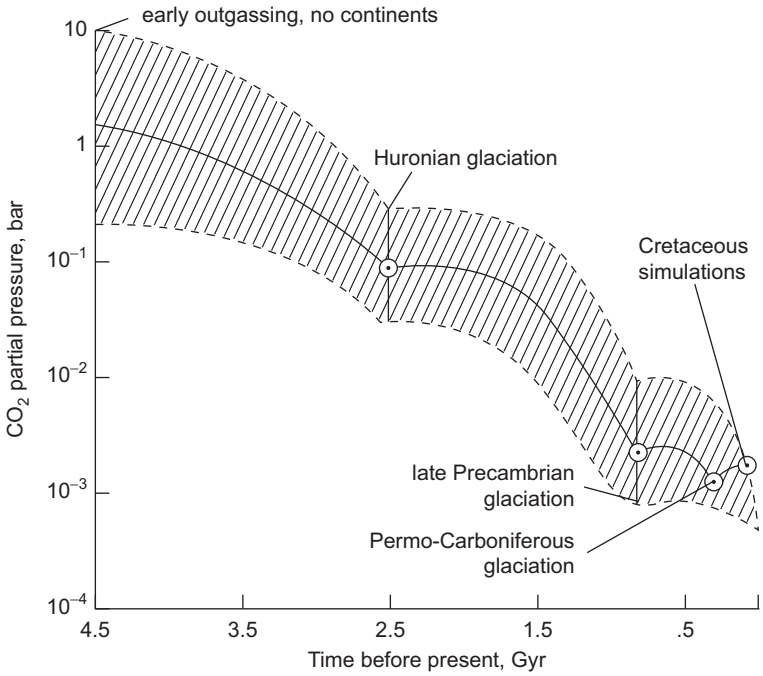


Figure 14.8 History of the CO_2 partial pressure in Earth's atmosphere based on various lines of evidence regarding Earth's surface temperature. Changes in slope at 2.5 and 1 billion years are notional only, based on occurrence of ice age episodes. "Cretaceous simulations" are described in Chapter 19. Adapted from Kasting and Ackerman (1986).

of carbon dioxide from the atmosphere. Sedimentation rates are higher, but the rates of subduction and carbon dioxide outgassing are not affected. As a result, carbon dioxide abundance in the atmosphere decreases, weakening the greenhouse effect and lowering temperature.

The carbon–silicate cycle possesses what is called a *negative feedback* loop in which changing the conditions in one direction tends to cause the system to move back in the other direction. This is characteristic of physical systems that are stable and provides at least a partial explanation as to why Earth's climate has remained in the temperature range allowing liquid oceans over geologic time. Absent such a feedback, changes in the Sun's luminosity, in the rate of spreading of plates, and in the amount of volcanism, as well as disasters such as giant impacts (Chapter 18) could have moved Earth's climate out of the range in which life is sustainable.

Life itself has played a role in altering carbon loss rates: the development of soil-forming microorganisms some 3 billion years ago accelerated the trapping of carbon dioxide in soils and hence may have led to a net decrease in carbon dioxide levels in the late Archean atmosphere. The development of calcareous plankton shifted most of the deposition of carbonate to deep ocean rather than to shallow, continental-shelf environments, hastening the transport of carbonates to subduction zones and increasing the rate of reintroduction of carbon dioxide to the atmosphere. These and other evolutionary changes in Earth's biosphere have thus caused shifts in climate to which life has had to adjust through the formation of new species (Chapter 18). The British scientist James Lovelock, and others, have even proposed that life acts to control the environmental feedback processes to maximize the habitability of Earth. This controversial "Gaia" hypothesis is thought provoking but not required necessarily to explain the stability of Earth's climate. Perhaps instead, life is a somewhat meddlesome passenger largely along for the ride.

14.7.3 The carbon–silicate cycle during the Archean

Although the carbon–silicate cycle seems to be a good candidate for explaining how Earth's climate has been stabilized over time, it is necessary to think carefully about how it operated during the Archean, when conditions were different from today. Smaller amounts of continental mass exposed above sea level could have reduced the efficacy of silicate weathering, which is the step that determines the rate of carbon dioxide sequestration. More rapid recycling of crust at the time would have led to a faster return of carbon dioxide to the atmosphere, with less stored in sediments on the seafloor. Both of these somewhat speculative differences between the Archean and more modern times mitigate in favor of leaving the available carbon dioxide in the atmosphere, rather than locked in sediments, during the Archean.

The high carbon dioxide abundances required to sustain Archean temperatures above that of the water freezing point led to a potential instability in the early Archean: the Sun's luminosity was low enough that the feedbacks in the carbon–silicate cycle might not have worked to bring Earth out of an ice-covered state, if it fell into one during that time. The reason for this lies

in a phenomenon that we discuss for Mars in Chapter 15 – cold temperatures and high carbon dioxide abundances would have caused carbon dioxide clouds to form, with the cooling effect of these clouds short-circuiting the gas's ability to warm the surface. This is a problem unique to the Archean because, for higher solar luminosities, less carbon dioxide is required to drive Earth out of a global ice age, so that carbon dioxide cloud formation is not an issue. However, some models suggest that carbon dioxide clouds might, under certain conditions, actually warm the surface (Chapter 15).

14.8 A balance unique to Earth, and a lingering conundrum

The history of Earth's atmosphere has been one of declining carbon dioxide abundance from the Archean onward. Geologic evidence for extensive or near-global ice coverage – "glaciation" – in the middle Archean (2.9–2.7 billion years ago), and the late Archean/early Proterozoic (2.4–2.2 billion years ago), and the lack of iron carbonates in paleosols are important constraints, as is glaciation in the late Precambrian (Figure 14.8). The decline has been slow because of the buffering effects of the carbon–silicate cycle, operating on a planet with liquid water and plate tectonics. The two Archean ice ages might, speculatively, have been triggered by rises in oxygen production due to photosynthesis (Chapter 17), which would have rapidly destroyed essentially all the methane in the atmosphere (and, by thus removing any protective organic haze, the ammonia too), thereby cooling the Earth. Or one or both might have been caused by the progressive decline in carbon dioxide abundance. Frustratingly, because of the paucity of the rock record throughout the Archean, we may never know.

If Earth did not possess liquid water, erosion of rock would be extraordinarily slow. On Venus, the surface pressure is 90 atmospheres of carbon dioxide. The total amount of carbon dioxide in Venus' atmosphere is not very different from the total equivalent of carbon dioxide locked up in various forms in Earth's crust. As discussed in Chapter 15, Venus lost whatever water it had early in its history, forcing all of the carbon dioxide to remain in the atmosphere. The resulting massive greenhouse effect has pushed the surface temperature of Venus to 730 K, much too high for liquid water to exist. Venus is trapped outside the carbonate cycle with no way to rid itself of the carbon dioxide that keeps the surface so warm.

Mars has a thin atmosphere, and its surface is too cold to support an ocean. Evidence (Chapter 15) shows that early Mars had episodes in which liquid water existed on its surface, but the crust of Mars shows no evidence of plate recycling – it is a small planet with little internal heat and hence lethargic tectonics. No recycling means that the carbonates formed from carbon dioxide essentially will never yield the greenhouse gas again. The liquid water on early Mars would have encouraged weathering, carbonate formation, and decreasing carbon dioxide – until the temperature of the surface got so low that the water froze and the weathering process ground to a halt.

In this "Goldilocks" view, Earth has two unique things that keep the cycle going: (i) liquid water to make carbonates from carbon dioxide, and (ii) a vigorous recycling of the ocean crust, which releases much of the carbon dioxide back into the gas phase. Earth is far enough from the Sun to enable liquid water to exist without catastrophically vaporizing and escaping, the fate of Venus discussed in Chapter 15. Earth also is much larger than Mars, which is too small to have the tectonic recycling needed to keep carbon dioxide in the atmosphere. Nonetheless, the atmospheric supply of carbon dioxide on Earth has continued to decrease slowly over time, from as much as 10 atmospheres in the Hadean or Archean, through to the present value of 0.0003 atmospheres. And, in spite of the increasing output of the Sun, Earth's climate generally has gotten cooler over time: the first glacial episode for which evidence exists was just after the close of the Archean. In these cold glacial episodes, less rainfall has meant that less carbon dioxide is lost from the atmosphere, allowing build up of the gas to warm the climate again and encouraging faster subsequent loss of carbon dioxide from the atmosphere. Therefore, the overall march toward less CO_2 and cooler climes appears to be a theme of Earth history up to the present.

Although the story presented above seems tidy, not all issues are resolved. It remains a serious problem that carbon dioxide alone may not have been enough to explain warm temperatures on Earth in the weak glow of the faint early Sun. Explaining warm episodes on Mars, as we show in Chapter 15, exacerbates the problem. Louisiana astronomer D. Whitmire and colleagues have proposed a quite different solution: that the early Sun was not low in luminosity after all. To get around the seemingly simple solar physics that led to the faint-sun problem, they suggest that the Sun was more massive 4 billion years ago and has lost that mass through expulsion of hydrogen gas in the form of a wind.

The present solar wind, a tenuous medium of protons and other ions, is much too weak to remove the mass required to create a brighter early sun, and so, the hypothesis must rely on the notion that the mass loss was much higher earlier in the history of the solar system. The viability of the idea rests on observing such mass loss from stars similar in age to the Archean Sun elsewhere in the galaxy, an observation that currently is very difficult. Nonetheless, the proposal itself reminds us that the keys to understanding the history of Earth lie buried not only in our own planet, but in our planetary neighbors, in the Sun, and in the neighboring galactic regions illuminated by the burning of billions of other suns.

Summary

Earth during the Archean possessed liquid water on its surface, a situation no different from that of today. However, the basic physics of nuclear fusion dictates that the Sun was 25% less luminous 3.8 billion year ago than it is today, and hence if all else were the same, the oceans of the Earth should have been frozen over. A number of solutions to this problem have been proposed, including the possibility that the Sun was more massive (implausible), or that Earth's atmosphere had a larger quantity of "greenhouse gases" at the time. Greenhouse gases refer to infrared absorbing molecules present in an atmosphere that is more transparent in the optical part of the spectrum than in the infrared. Sunlight arriving at the Earth is greatly diluted in the number of photons per unit area relative to what was emitted at the surface of the Sun. As the photons are absorbed by the ground, they are re-emitted as a much larger quantity of infrared photons corresponding to a lower black-body temperature – a consequence of the second law of thermodynamics. These infrared photons are impeded in their movement outward through the atmosphere as they are absorbed and re-emitted by greenhouse gases. Since the energy coming in per second must balance the energy going out per second, the atmosphere responds to this imbalance between the transparency in the optical and opaqueness in the infrared by making the temperature profile steeper. Hence the surface temperature is elevated relative to the case of no atmosphere or a fully transparent atmosphere. The primary greenhouse gases today in Earth's atmosphere are water, carbon dioxide, and methane. Water is controlled by evaporation and condensation; carbon dioxide is a small fraction of the total carbon that may have existed as carbon dioxide in the past, and so the early Earth's atmosphere could have had more carbon dioxide to compensate for the fainter Sun. However, minerals that should be in the rock record if CO_2 had been vastly more abundant are absent, and this may limit the amount of carbon dioxide that can be invoked for the Archean Earth. Instead, other greenhouse gases such as methane might have played an important role. Complicating this is the role of clouds, which can cool or warm the surface depending on the altitude at which they form. The geologic record suggests that at times in the Archean and subsequent eons, the Earth plunged into deep ice ages, indicating either a quite variable Sun, or fluctuations in the amount of greenhouse gas present in the atmosphere over time. Carbon dioxide would have gradually been scrubbed from the atmosphere by the carbon–silicate cycle, ending up as carbonates on the seafloor. However, plate tectonics recycles some of the carbonates back into carbon dioxide, an essential recharging mechanism for the atmosphere without which the Earth might have been much colder throughout its history. Mars gives us an example of a planet that likely had a thick greenhouse atmosphere early in its history, which it then lost along with its surface environment capable of sustaining liquid water in a stable fashion.

Questions

1. The presence of carbon recycling on Earth, as a buffer against the faint early Sun, and excessive temperatures later, might strike some as a kind of "just right" story, such that few planets other than twins of Earth could sustain life. What other kinds of processes could keep a climate habitable for life?
2. Is there any limit on planets much more massive than Earth sustaining life? Could a rocky body 10 times Earth's mass sustain life? What might be the problems?
3. What is the dilution factor between the number of photons per unit area at the surface of the Sun and at a shell corresponding to the radius of the Earth's orbit? How would you then calculate the equilibrium black-body temperature corresponding to the energy received per unit area and per time at that distance from the Sun?
4. Do a literature search on the evidence for and against the faint young Sun. How plausible is the possibility that the Sun has lost mass with time? How much mass would it need to lose?
5. What are some of the complexities associated with trying to reconstruct the past atmospheric composition in the Archean?

General reading

Kasting J. and Catling, D. 2003. Evolution of a habitable planet. *Annual Review of Astronics and Astrophysomy* **41**, 429–63.

Williams, G. R. 1996. *The Molecular Biology of Gaia*. Columbia University Press, New York.

References

Falkowski, P., Scholes, R. J., Boyle, E. *et al.* 2000. The global carbon cycle: a test of our knowledge of Earth as a system, *Science* **290**, 291–6.

Haqq-Misra J. D., Domagal-Goldman S. D., Kasting P. J., and Kasting J. F. 2008. A revised, hazy methane greenhouse for the early Earth. *Astrobiology* **8**, 1127–37.

Houghton, J. T. 1977. *The Physics of Atmospheres*, 1st edn. Cambridge University Press, Cambridge, UK.

Kasting, J. F. 1989. Long-term stability of the Earth's climate. *Paleogeography, Paleoclimatology, Paleoecology* **75**, 83–95.

Kasting, J. F., and Ackeman, T. P. 1986. Climatic consequences of very high CO_2 levels in the Earth's early atmosphere. *Science* **234**, 1383–5.

Kharecha, P., Kasting, J. F., and Siefert, J. L. 2005. A coupled atmosphere–ecosystem model of the early Archean Earth. *Geobiology* **3**, 53–76.

Knauth, L. P. 1992. Origin and diagenesis of cherts: an isotopic persective. In *Isotopic Signatures and Sedimentary Records* (N. Clauer and S. Chandhuri, eds). Springer-Verlag, Berlin, pp. 123–52.

Nutman, A. P., Mojzsis, S. J., and Friend, C. R. L. 1997. Recognition of ≥3850 Ma water-lain sediments in Greenland and their significance for the early Archean Earth. *Geochimica Cosmochimica Acta* **61**, 2475–84.

Pavlov, A. A., Hurtgen, M. T., Kasting, J. F., and Arthur, M. A. 2003. Methane-rich proterozoic atmosphere? *Geology* **31**, 87–90.

Peixoto, J. P. and Oort, A. H. 1992. *Physics of Climate*. AIP Press, New York.

Rosing, T., Bird, D. K., Sleep, N. H., and Bjerrum, C. L. 2010. No climate paradox under the faint early Sun. *Nature* **464**, 744–7.

Sagan, C. and Chyba, C. 1997. The early faint Sun paradox: organic shielding of ultraviolet-labile greenhouse gases. *Science* **276**, 1217–21.

Sheldon N. D. 2006. Precambrian paleosols and atmospheric CO_2 levels. *Precambrian Research* **147**, 148–55.

Trenberth, K. E., Houghton, J. T., and Meira Filho, L. G. 1996. The climate system: an overview. In *Climate Change 1995: The Science of Climate Change* (J. T. Houghton, L. G. Meira Filho, B. A. Callander, N. Harris, A. Kattenberg, and K. Maskell, eds). Cambridge University Press, Cambridge, UK, pp. 51–65.

Valley, J. W., Peck, W. H., King, E. M., and Wilde, S. A. 2002. A cool early Earth. *Geology* **30**, 351–4.

Whitmire, D. P., Doyle, L. R., Reynolds, R. T., and Matese, J. J. 1995. A slightly more massive young sun as an explanation for warm temperatures on early Mars. *Journal of Geophysical Research* **100**, 5457–64.

Wolf E. T. and Toon O. B. 2010. Fractal organic hazes provided an ultraviolet shield for early Earth. *Science* **328**, 1266–68.

Young, G. M., von Brunn, V., Gold D. J. C., and Minter, W. E. L. 1998. Earth's oldest reported glaciation; physical and chemical evidence from the Archean Mozaan Group (2.9 Ga) of South Africa. *Journal of Geology* **106**, 523–38.

15

Climate histories of Mars and Venus, and the habitability of planets

Introduction

Earth at the close of the Archean, 2.5 billion years ago, was a world in which life had arisen and plate tectonics dominated the evolution of the crust and the recycling of volatiles. Yet oxygen (O_2) still was not prevalent in the atmosphere, which was richer in CO_2 than at present. In this last respect, Earth's atmosphere was somewhat like that of its neighbors, Mars and Venus, which today retain this more primitive kind of atmosphere.

Speculations on the nature of Mars and Venus were, prior to the space program, heavily influenced by Earth-centered biases and the poor quality of telescopic observations (Figure 15.1). Forty years of US and Soviet robotic missions to these two bodies changed that thinking drastically. The overall evolutions of Mars and Venus have been quite different from that of Earth, and very different from each other. The ability of the environment of a planet to veer in a completely different direction from that of its neighbors was not readily appreciated until the eternally hot greenhouse of Venus' surface and the cold desolation of the Martian climate were revealed by spacecraft instruments.

However, robotic missions also revealed evidence that Mars once had liquid water flowing on its surface. It is tempting, then, to assume that the early Martian climate was much warmer than it is at present, warm enough perhaps to initiate life on the surface of Mars. However, the difficulty of sustaining a warm Martian atmosphere in the face of the faint-early-sun problem of Chapter 14 remains a daunting puzzle, one that is highly relevant to the broader question of habitable planets beyond our solar system. What is the range of distances from any given star for which liquid water is stable on a planetary surface and life can gain a foothold?

In the temporal sequence that Part III of the book has been following, we stand near the end of the Archean eon. By this point in time, the evolution of Venus and its atmosphere almost certainly had diverged from that of Earth, and Mars was on its way to being a cold, dry world, if it had not already become one. This is the appropriate moment in geologic time, then, to consider how Earth's neighboring planets diverged so greatly in climate, and to ponder the implications for habitable planets throughout the cosmos. In the following chapter, we consider why Earth became dominated by plate tectonics, but Venus and Mars did not. Understanding this is part of the key to understanding Earth's clement climate as discussed in Chapter 14.

15.1 Venus

15.1.1 Origin of Venus' thick atmosphere

The atmosphere of Venus contains somewhat more nitrogen than does that of the Earth: 3 atmospheres of pressure instead of 0.8 atmospheres. More striking, however, is the enormous surface pressure of 90 atmospheres of carbon dioxide. The consequence of Venus' massive atmosphere is an enormous greenhouse effect: even though the clouds of Venus' upper atmosphere, largely sulfur compounds, reflect much more sunlight away than do the clouds of Earth, Venus has a surface temperature of 730 K. In other words, even though the surface of Venus receives less sunlight than does the Earth's surface, the temperature at Venus' surface is above the melting point of lead. Liquid water is not stable on the surface or anywhere in the atmosphere. Gaseous water vapor is only 20 parts per million by number – that is for every million carbon dioxide molecules, there are 20 molecules of water. Oxygen is not abundant either, with a pressure of 0.002 atmospheres, 1% that in our atmosphere.

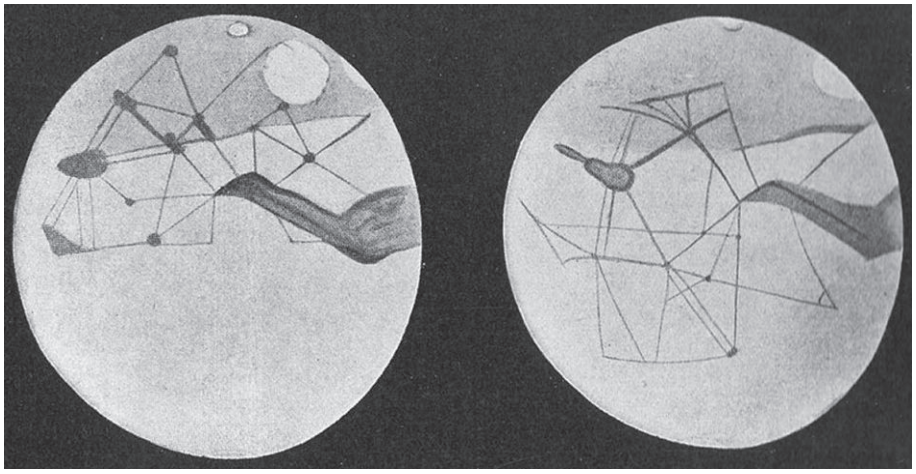


Figure 15.1 Prior to the use of photographic and electronic detectors, maps of Mars sketched by hand typically showed unnaturally straight lines, a result of atmospheric turbulence that blurred telescopic images and caused the merging of irregular dark features. Such lines were considered by some as a sign of intelligence. At the turn of the twentieth century, the American astronomer Percival Lowell interpreted these illusory features as vast canals bringing water from the Martian polar caps to the parched equatorial deserts, a grander version of what was actually undertaken at the time in the Arizona and California deserts south and west of his high plateau observatory.

How Venus came to this state is still a subject of heated debate. Venus is almost the same size as Earth, of similar density (and hence internal composition), and somewhat nearer to the Sun. One clue is the close correspondence of the amount of carbon dioxide in Venus' atmosphere with the amount of carbon dioxide that could be produced from the carbonates and other carbon compounds trapped today in Earth's crust. If Earth's oceans were to boil away, and the hydrological cycle of rainfall end, recycling of carbonates into the atmosphere might eventually build up a massive carbon dioxide atmosphere on our planet as well. The divergent evolutionary paths that Earth and Venus have taken apparently have to do with the lack, or early loss, of large quantities of water from Venus. Direct measurement of Venus' atmosphere from *Pioneer* Venus entry probes in 1978 revealed a large abundance of deuterium (defined in Chapter 2) relative to light hydrogen in the atmosphere of Venus, the ratio of the two being about 150 times that in the oceans of Earth. One interpretation of such an overabundance is that large amounts of water escaped from Venus early in its history; as the water was lost in gaseous form from the atmosphere, the heavier deuterium atoms in HDO and D_2O (versus H_2O) were more likely to be retained. Although alternative models have been proposed (for example, that the high deuterium abundance is a contaminant from impacting comets), the water-loss model appears at present to be the best explanation for the deuterium data.

If Venus did have liquid water early in the solar system's history, the challenge is to understand how it was lost and when. The traditional explanation for the loss lies in the so-called *runaway greenhouse*, featured in many textbooks. Here, the solar heating at Venus' distance from the Sun, coupled with a sufficient amount of initial greenhouse heating from water and carbon dioxide, leads to an unstable situation: heating causes more evaporation of water from the ocean (because the evaporation rate and the total water vapor content in the atmosphere are very sensitive to the temperature). This higher water content, in turn, increases the atmospheric temperature through the greenhouse effect, which in turn causes more water to evaporate, warming the atmosphere further. The system enters a "runaway," leading quickly to the complete boiling away of the oceans.

Very careful modeling of the early history of Venus shows that at the time, a runaway greenhouse was marginal for that planet. The reason lies again in the faint-early-sun problem. Although today Venus receives 1.9 times the amount of sunlight that Earth does at the top of the atmosphere (remember much of this is reflected by Venus' clouds), in the earliest period of solar system history the sunlight that Venus received was only 1.4 times that received by Earth at present. Below a certain threshold surface temperature, the greenhouse effect does not evaporate enough water to initiate a runaway.

So how did Venus arrive at its present state? The solution to this puzzle lies in considering the effect of water vapor on the entire atmosphere, as shown in Figure 15.2. On Earth today, because the temperature drops rapidly with altitude as the atmosphere thins and becomes more transparent to infrared radiation, the amount of water vapor drops very steeply. At about 10 km above the surface lies a boundary between the lower atmosphere, the *troposphere*, and the *stratosphere* above it. This boundary, the *tropopause*, is defined by the altitude at which the temperature stops falling and begins rising at higher altitude as the air becomes transparent to most infrared radiation, and some molecules selectively absorb sunlight in the ultraviolet wavelengths. Above the tropopause, water vapor no longer decreases with increasing altitude; its minimum value is determined by the temperature at the tropopause.

In Earth's atmosphere today, the falloff of temperature with height leads to a very sharp decline in water vapor with altitude. The water vapor condenses as clouds and these eventually are lost as rain. The Earth's stratosphere is extremely dry today, about as dry as the present bulk atmosphere of Venus. What water vapor does exist in the stratosphere is subject to being broken apart by ultraviolet photons from the Sun to form oxygen (O_2) and hydrogen; because hydrogen is a light molecule,

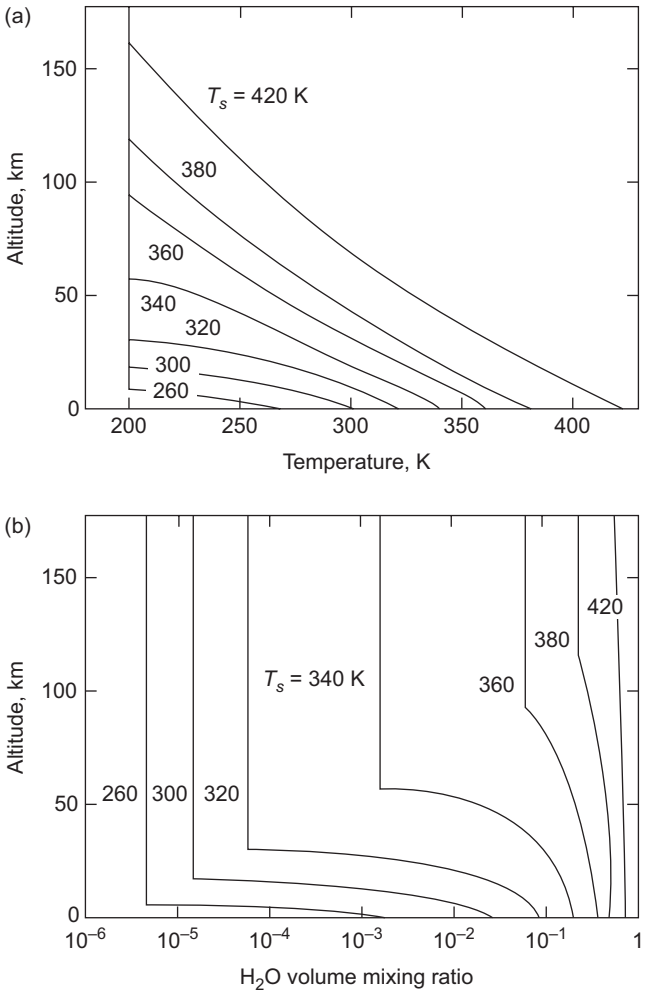


Figure 15.2 A moist greenhouse atmosphere in action. The temperature (a) and amount of water vapor (b) are plotted versus altitude for different values of the surface temperature. Each profile is marked with its particular surface temperature, T_s. The water volume mixing ratio is simply the number of water molecules divided by the number of all molecules (of all chemical species) in the atmosphere at a given altitude. Hence a water mixing ratio of 10^{-3} means that one out of every thousand molecules is water. The stratosphere is simplified in the calculation by assuming that it has a constant temperature of 200 K; in reality, its temperature is not constant. See text for a description of the moist greenhouse loss of water. Reproduced from Kasting (1988) by permission of Academic Press.

it moves upward in the atmosphere and eventually is lost to space. The ultraviolet radiation is restricted to high altitudes precisely because it is absorbed there by molecules such as water and ozone; the vast majority of Earth's water is protected from such destruction by being resident in the oceans and lower atmosphere.

Consider now what would happen if Earth's surface temperature were increased, simulating what might have happened on Venus if it once had had liquid water oceans. More water vapor is put into the troposphere, allowing formation of more massive cloud decks. Clouds can warm or cool the climate, depending on their altitude, but their formation by condensation always releases heat, which causes the temperature profile to fall more gently with altitude. Because of this effect, the temperature profile for higher surface temperatures declines more gradually than for lower surface temperatures, and the tropopause boundary between the troposphere and the stratosphere shifts upward as the surface warms (Figure 15.2). More water is admitted into the stratosphere, and eventually large amounts of water are present at altitudes accessible to solar ultraviolet photons. For a surface temperature just 80 K above Earth's current global mean value, the water vapor at high altitudes increases by a factor of 10,000.

In effect, then, a global surface temperature above 340 K "pops the cork" on the water budget of the atmosphere, allowing large amounts of water vapor to flow to altitudes where solar ultraviolet radiation breaks it apart, and the hydrogen escapes. This *moist* greenhouse crisis operates at lower solar fluxes than is required for the runaway greenhouse; for an Earth-like atmosphere with nitrogen and a small amount of CO_2, the threshold for the moist greenhouse may be as low as 1.1 times the present solar flux received by Earth. This flux is well below that which was received by Venus during the faint-early-sun epoch, but above that for Earth throughout its history.

We can imagine what happened to Venus early in its history. Possessed of an atmosphere with at least as much CO_2 in gaseous form as Earth possesses today, but lacking the present-day global layer of sulfurous clouds that reflects much of the sunlight away, Venus' surface was above the temperature threshold for the moist greenhouse even when the solar flux was only 70% of its present value. If liquid water did exist on the surface at the time, the atmospheric temperatures were high enough to allow evaporated water to flow freely to the tenuous upper atmosphere. Ultraviolet photons broke up the water molecules, causing most of the hydrogen to be lost and eventually depleting the planet of water. The signature of this lost ocean is with us today in the form of a high Venusian ratio of deuterium to hydrogen, because the heavier deuterium tended to be left behind in the atmosphere as hydrogen escaped.

Once bereft of surface water, the die was cast for Venus. Carbon dioxide in the atmosphere had no means of being locked up in surface rocks because liquid water was not available to efficiently make hydrogen carbonates. The carbon dioxide that we see today in Venus' atmosphere escapes only very slowly from the top of the atmosphere, cannot be trapped in rocks at the surface, and thus remains as a massive gaseous memento of the early loss of water.

How quickly could the water have been lost? Observations of young stars suggest that the early Sun put out more ultraviolet radiation than it does today, though, as discussed in Chapter 14, its overall energy output was lower. Based on the amount of ultraviolet radiation available at Venus from the early Sun, less than 100 million years were needed to remove the equivalent of an Earth's ocean-worth of water. Hence there was little or no time to lock up carbon dioxide as carbonates before the water was lost, and because accretional heat likely was still contributing to a very hot early crust for Venus, most or all of Venus' carbon dioxide complement was likely *never* locked up in the crust. The massive amount of CO_2 present today in the atmosphere is probably close to the original atmospheric abundance, although some of the carbon dioxide could have been added later from the Venusian mantle by volcanoes. An alternative view is driven by the possibility that the moist greenhouse runaway does not

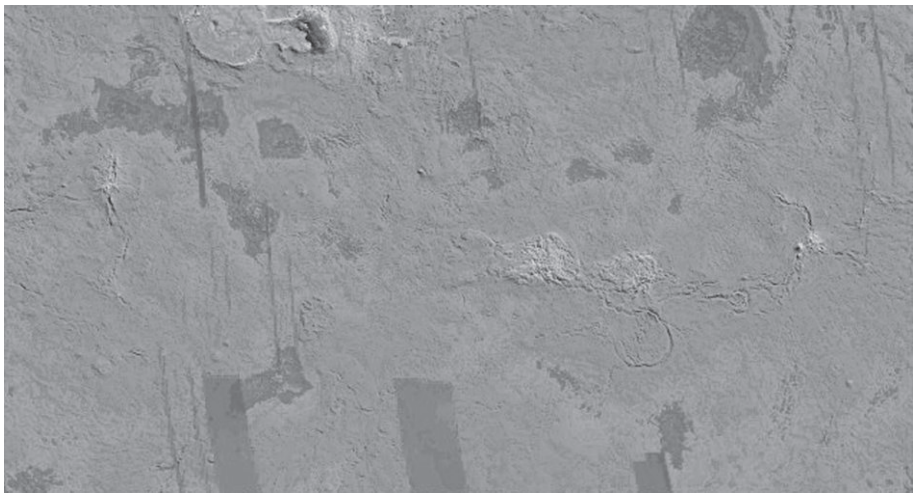


Figure 15.3 Global topography of Venus. Red areas are highest, blue lowest. Courtesy of NASA/Jet Propulsion Laboratory. See color version in plates section.

operate as efficiently as described above, requiring more sunlight before it is triggered, or that Venus did indeed have a layer of clouds early on that obscured the surface. In this view, the runaway that removed essentially all of Venus' water did occur, but later in the planet's history, when the solar luminosity had grown sufficiently. A late loss of Venus' surface oceans might have allowed life to form, and then perish as the habitable environment was lost forever. It also would have allowed rocks to be transformed by the presence of water in the Venusian crust, producing hydrated basalts, andesites, or even granites. A future mission to search for such types of rocks, perhaps exposed in the more mountainous terrains of the Venusian surface, could test whether Venus' transformation into hell occurred after a substantial period of habitability. The moist greenhouse model has important consequences for the habitability of planets in general, a point we return to in section 15.6.

15.1.2 Overview of the surface of Venus

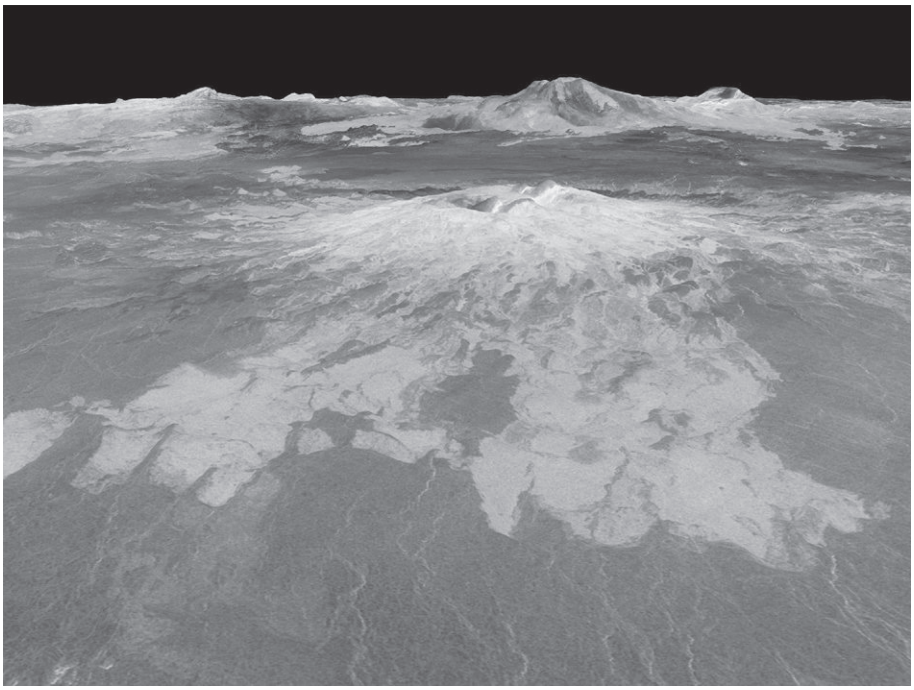
Although early Soviet and US probes measured the atmospheric composition and temperature of Venus, mapping the geology of the surface was hindered by a global mass of sulfurous clouds at high altitude. First radar from Earth, then radar from two Soviet orbiters *Veneras* 15 and 16 and the US *Magellan* spacecraft have enabled mapping of the surface. A radar mapper functions like a camera that provides its own flash or source of illumination. Photons at radio wavelengths (Chapter 3) can penetrate the clouds, and the radar transmits such photons to the ground surface. These are reflected and scattered, and some are received back at the radar antenna. By coding or shaping the transmitted pulse of photons, and taking advantage of the orbital motion of the spacecraft, the received photons can be arranged or mapped by computer into an image of the surface at radio wavelengths. For detailed geologic work, the very high resolution *Magellan* images, collected at Venus from 1990 through 1993, are of greatest use.

The geology of Venus, on a broad scale, looks at first glance like the Earth's with highlands rising out of a lowland plain, akin to continents rising above the ocean floor. However, the proportion of land on Venus that rises above the mean surface elevation is far smaller than on Earth; likewise, there are few long, deep cuts in the crust like Earth's submarine trenches (Figure 15.3). Thus the signatures of mature plate tectonics – massive continents and subduction zone trenches – are largely missing. It is as if we were to look at Earth in the Archean eon of time, when plate tectonics was just getting going and continental masses were small. Soviet probes have sampled several regions on the surface; all of the analyses are consistent with basaltic compositions (close to that of Earth-ocean crust), but the accuracy of the technique, and regions covered, were limited.

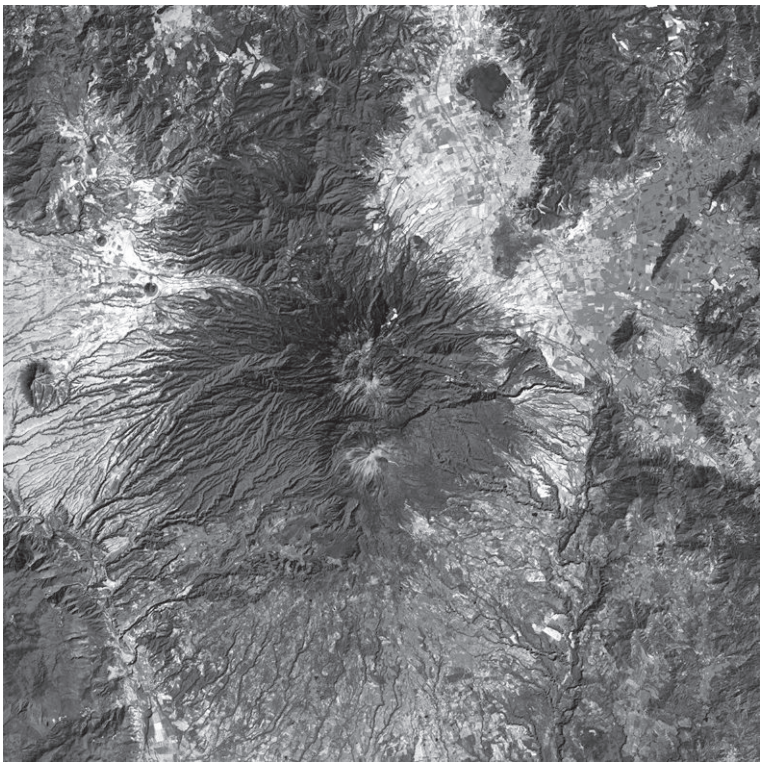
The surface of Venus contains impact craters. Although the number of these is far larger than on Earth, it is smaller than that of the Moon and Mars. The number of craters is consistent with a surface that has renewed itself through volcanic flows over geologic time, with the last overall renewal of the surface being perhaps 300 million to 600 million years ago. (Whether the surface is continually or episodically active geologically is addressed in Chapter 16.) This is long after the loss of any putative Venusian water ocean, even if the latter occurred after billions of years of Venusian history. Hence, any evidence of ancient oceans is mostly buried under the late volcanic veneer, with the possibility that some outcrops of the original crust are exposed in places.

The thick atmosphere prevents small bolides from reaching the surface; however, the largest impactors smack into the surface at high velocities, unimpeded by the atmosphere. Because there is no surface water on Venus, craters and other landforms that are not buried in lava erode very slowly. The mean slope of features is therefore larger on Venus than on Earth, and the images of mountain ranges are eerie in their evident absence of water erosion (Figure 15.4).

The apparent lack of plate tectonics and its accompanying geologic signatures on Venus is perhaps the most profound difference between Venus and Earth. Remarkably, the presence of water is apparently an important condition for sustaining plate motions, and certainly for the formation of continental masses



(a)



(b)

Figure 15.4 (a) Sapas Mons, a 600 km diameter, 1.5 km high volcano on Venus, shows no evidence of water erosion; the bright linear features have the form and appearance of lava channels. This *Magellan* radar view exaggerates the vertical extent by a factor of 10. Image courtesy of NASA/Jet Propulsion Laboratory. (b) Snow-capped Colima Volcano in Mexico. The southern caldera has been active historically. Calderas and flanks show an intricate network of water-carved channels. The image was made by the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) aboard NASA's Terra satellite. Courtesy of USGS. See color versions in plates section.

that are present on Earth in abundance (and may only exist on Venus in one location, if at all). We defer a more detailed development of this idea to Chapter 16, where the origin of Earth's plate tectonic geology is explored and compared to Venusian geology. Significant and striking geologic differences are apparent on these two planets that should be ridding themselves of the same amount of internal heat; understanding the origin of these differences is perhaps the most important question in Venusian geology.

15.2 Mars

15.2.1 Mars today

The Martian atmosphere is, in composition, very similar to that of Venus, with carbon dioxide most abundant, nitrogen the secondary constituent, and water and oxygen in minor abundance. Mars' atmosphere is diminutive compared to those of Venus and Earth, however. The surface pressure is only 0.006 of an atmosphere. The thin atmosphere means that Mars has hardly any greenhouse warming. This, combined with its greater distance from the Sun, results in a temperature range from as much as 270 K at the equator to only 150 K at the polar caps. Mars is a true opposite of Venus: a cold dry planet, with air so thin that ultraviolet rays from the Sun penetrate to the surface, effectively sterilizing its uppermost soil.

Mars is so cold that the carbon dioxide atmosphere freezes out seasonally at the poles. The pressure in the atmosphere therefore varies significantly over the Martian year, which is about twice an Earth year. The tilt (*obliquity*) of Mars currently is the same as Earth's; the summer sun shines on one pole, evaporating carbon dioxide and driving it to the winter pole. Mars' axis, however, may undergo large shifts in its obliquity caused by gravitational tugging of the other planets, principally Jupiter; Earth would suffer the same fate were it not for the stabilizing effects of our large Moon. There is some faint evidence in geological features across the Martian surface that past tilt may have exceeded 50 degrees (the current value is 24 degrees).

About one year out of two, heating during the southern hemisphere spring drives large quantities of dust into the atmosphere, allowing more sunlight to be absorbed in the atmosphere and moving dust across the planet. These global dust storms may last for weeks or months.

Water is present today on Mars as ice trapped at one or both polar caps, but probably is more abundant as ground ice trapped in a zone of permanent freezing (*permafrost*) throughout high- and mid-latitude regions of Mars' crust. Water ice also condenses out in the thin atmosphere; storm systems occasionally have been seen in orbiting spacecraft images. The search for life on Mars began with the landing of two sophisticated robot laboratories, *Vikings* 1 and 2, in 1976. These laboratories sampled Martian soil and tested for chemical reactions that might indicate living processes. No evidence of life was found in the dry regions to which the landers had been targeted: sites that were chosen to maximize the chances of safe landings. Furthermore, the abundance of organic molecules on the surface was so low as to be undetectable. The thin atmosphere of Mars, with no ozone shield, allows solar ultraviolet radiation to penetrate to the surface and break apart chemical bonds; organic molecules are readily destroyed in such an environment, and much of the hydrogen is lost to space. Additionally, the iron in the soil is combined with oxygen in such a way as to make an extremely reactive mixture that would quickly oxidize organic molecules. The present surface of Mars, at least in the high plains, is an inhospitable location for life.

15.2.2 Martian geology

Unlike Venus, the Mars surface is visible at all times except during dust storms. Cameras sent to Mars on robotic missions have mapped the surface in great detail from orbit and at three landing sites. The geology of Mars is very different from that of the Earth, in that the Martian crust is not being shifted around on plates nor recycled in the interior. Magma brought to the surface continues to pile up into enormous volcanoes dwarfing any on the Earth – a paradox, since Mars appears to be much less active at present than is the Earth. An enormous canyon system, Valles Marineris, adjacent to the giant volcanic shield of Tharsis Mons, represents such a dramatic and singular crustal rupture that it speaks to the idea that individual pieces of crust cannot move anywhere.

On the large scale, Mars shows no evidence for continents and lowland (ocean-type) basins. The southern hemisphere stands several kilometers above the northern hemisphere; it is dominated by heavily cratered highlands while the northern plains are smooth, suggesting a blanket of either volcanic debris or sediments from a past ocean. This asymmetry may be the result of a giant impact early in Mars' history; it is not at all what one would get with Earth-style plate tectonics. Two extensive uplands on Mars are sites of past volcanism. The largest one, Tharsis, contains huge shield volcanoes, giant versions of the Hawaiian volcanoes. Again, they are clues to the static nature of the crust: with no lateral movement, magma welling up from the interior keeps spewing out material on the same part of the nonmoving crust, building up huge volcanoes in isolated locations. The *Viking* robotic landers sampled the soils at two widely separated landing sites in the northern hemisphere and found the rocks to be basaltic in composition. The Mars Exploration Rovers *Spirit* and *Opportunity*, which arrived on Mars in 2004, also found the rocks and dust around their landing sites to be basaltic in composition. The *Pathfinder* lander, arriving in Ares Valles in July 1997, identified one rock with an elemental composition consistent with andesite, which would be suggestive of plate tectonics. However, because only the elemental abundances could be determined on that mission, and not how the atoms are structured in a mineral, the finding was ambiguous: the rock could be an amalgam of basaltic material and more silica-rich debris from an impact.

The apparent lack of plate tectonics on Mars is almost certainly the result of its small size, but in a way that may seem counterintuitive. The small size of Mars allowed it to lose heat much more quickly than did Earth or Venus, and hence to form early in its history a crust much thicker than that of Earth. However, this thick crust actually impedes the transfer of heat to the surface of Mars, and then to space, because it is rigid and cannot convect. The result is that the temperature of the Martian interior may be too *high* for plane tectonics to operate, rather than

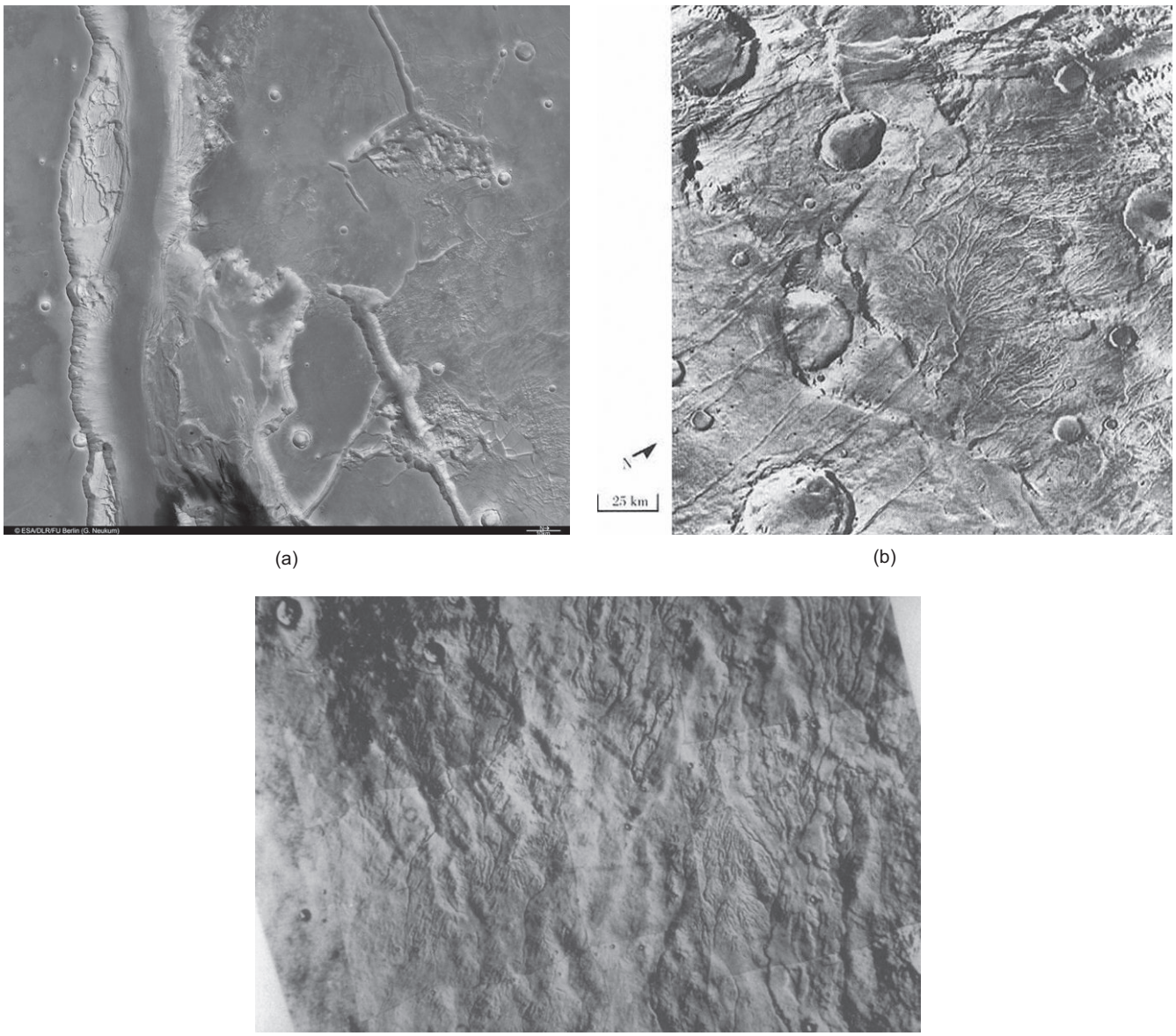


Figure 15.5 Three types of dried-up channels on Mars, in *Viking* Orbiter images: (a) Portion of an outflow channel (Kasei Valles). (b) Valley networks in the southern highlands of Mars. (c) Runoff channels on the volcano Alba Patera. Courtesy of NASA/Jet Propulsion Laboratory.

too low, at least according to some models. And it is certainly the case that a thick crust is difficult to fracture and to bend, essential properties for sustaining subduction zones. If there had been plate tectonics, it must have been a very early episode, and indeed faint magnetic traces suggestive of parallel strips arranged symmetrically around a line of symmetry have been mapped on the Martian surface by the *Mars Surveyor* operating in Mars orbit from 1997 to 2006. This suggests, albeit weakly, the possibility of spreading centers on ancient Mars akin to the mid-ocean ridges on Earth discussed in Chapter 9.

15.2.3 Geological hints of a warmer early Mars

Evidence for water on Mars abounds. Impact craters appear to have melted ground ice; their peripheries show signs of extensive mudflow. Volcanoes heat the ground and release water; a number of runoff channels reveal that water was melted by the eruptive heat. Most intriguing is evidence for a sustained earlier warm period on Mars contained in channels, canyons, surface deposits of carbonates and sulfates, and presence of evaporites and other minerals associated with liquid water:

1. Networks of dry channels and valleys are present on Mars. Three basic forms can be identified (Figure 15.5): *outflow channels*, *valley networks*, and *runoff channels*. The outflow channels appear to have been formed by the very rapid release of large quantities of water, or might have been carved by flows of debris (rocks, mud) mobilized by water. The flows in such channels were sufficiently energetic that they could have been sustained under virtually any atmospheric

conditions, including the cold, dry climate existing now on Mars (under which slowly flowing water would quickly freeze and then sublime to water vapor). The wide variation in abundance of craters on surfaces in and around the channels suggests that the channels formed episodically over the history of Mars. The valley networks, on the other hand, have a form that indicates they were carved by slowly flowing liquid water or, alternatively, by collapse of the surface (*sapping*) caused by groundwater flow. The possible sources of the water include melting of buried ice and expulsion to the surface, melting of surface ices, or even precipitation of snow or rain. The valley networks occur primarily, but not entirely, on surfaces that are very heavily cratered, and some of the impacts clearly occurred after the networks were formed. Most are therefore very ancient – dating to the end of the Heavy Bombardment some 3.8 billion years ago. Because their formation requires conditions very different from those present today (much more restrictive than required for the outflow channels), they could be a record of a time when the atmosphere was thicker and the climate warmer. A few younger valley networks, as well as runoff channels seen on the slopes of some volcanoes, suggest that warm conditions (possibly localized) may have occurred multiple times in Martian history.

2. Massive canyon systems formed by geologic processes show evidence of modification by liquid water. The canyons merge into numerous channels that show features caused by the flow of liquid water. Sedimentary deposits within the canyons have been seen on orbiting spacecraft images, which suggest the former presence of standing lakes.
3. The Martian surface contains exposures of carbonate and sulfate minerals, which usually if not always require water for their formation, as well as "phyllosilicates" – clays and other minerals formed in association with water. These were discovered and mapped by spectrometers on orbiting spacecraft such as *Mars Express* (2003–present), *Mars Reconnaissance Orbiter* (2006–present), and direct analysis of rocks examined by the Mars Exploration Rovers. *Spirit* found minerals consistent with water interacting with magma near the Martian surface; *Opportunity* found minerals such as hematite, jarosite, and others, that suggested in total the ancient presence of standing liquid water at the site that evaporated away, leaving the minerals behind. The minerals seen from orbit and from the Martian surface are strong chemical evidence for the presence of water in the ancient past on and under the surface of Mars in many places. However, the geographic extent of carbonates seen from orbit is much less than what one would have expected had a vast ocean been present, one that might have spanned what is today the northern hemisphere basin. The sulfates are more abundant and suggest that the ocean, if it existed, was quite acidic in composition – very different from Earth's ocean. The clays, though, speak to water with more neutral pH (neither acid nor alkaline), and so perhaps surface bodies of liquid water on Mars varied in their acidity with location, or time, or both. In any event, the simplest interpretation of the geochemical evidence is that liquid water was present on or near the surface of Mars for long periods, perhaps hundreds of millions of years, early in Martian history.
4. Some geologic features in various areas of Mars appear to have been carved by *glacial* action, that is, the movement of massive amounts of surface ices under their own weight. The features include certain kinds of ridges and troughs that resemble terrestrial landforms carved by glaciers and called *moraines* and *eskers*, as well as polygonal cracks typical of glacial terrains, and even lobate flows suggestive of debris-covered piedmont glaciers (Figure 15.6). If the glacial interpretation, first championed quantitatively by Jeffrey Kargel at the University of Arizona, is correct, it implies surface conditions in which water ice was stable against rapid sublimation, and hence requires conditions in which the atmosphere was denser than at present.
5. Water is present as ice in the Martian polar regions, and liquid water can appear when salts absorb water locally from the atmosphere – one interpretation of apparent droplets on the leg strut of the *Mars Phoenix* lander sitting in the high northern latitudes during Martian northern summer in 2008. But potentially vaster deposits of water are present beneath the surface at gradual increasing depths as one moves from the polar regions to the midlatitudes, as demonstrated by an Italian-built radar orbiting Mars aboard the US *Mars Reconnaissance* orbiter. A sister radar orbiting at longer wavelengths, aboard the European *Mars Express*, continues to probe for evidence of even deeper layers.

An early period of warm conditions on Mars, with liquid water, requires a thicker atmosphere of carbon dioxide, perhaps several atmospheres or more of pressure. Because it formed farther from the Sun than did Earth, in a cooler part of the solar nebula, Mars probably started out with at least as much water and carbon dioxide as did Earth. An early thick atmosphere is therefore possible. During such a period, life could have developed. Unlike on Earth, the climate apparently changed because carbon dioxide disappeared and temperatures fell below the freezing point of water, perhaps terminating Martian life. Whether warm conditions occurred in multiple episodes, and how recent the last such episode was, remain controversial. The interpretation of some Martian features as glacial in nature is an important part of the debate, because such features appear to be much younger than the bulk of the valley networks.

The cause of the climate cool down, or cool downs if there were multiple episodes of warmth, might be tied to Mars' small size. On early Mars, carbon dioxide could have been progressively locked up as carbonates in much the same way as on Earth (probably without the mediating step involving life). Mars, however, is much smaller than Earth and therefore has cooled more rapidly than our planet. The result seems likely to be a very thick crust that cannot slide horizontally in the form of recycling plates, as discussed above. Thus, on a Mars with no plate tectonic activity there was no means for significant recycling of the crust: carbon dioxide locked up as carbonates would have remained that way. Loss of atmosphere by impacts was also important, since the small size of Mars and hence weak gravity (one-third the Earth's) encouraged escape of gases heated by impactors. Whether carbonate formation or impact escape was the more important loss process is a matter of current debate.

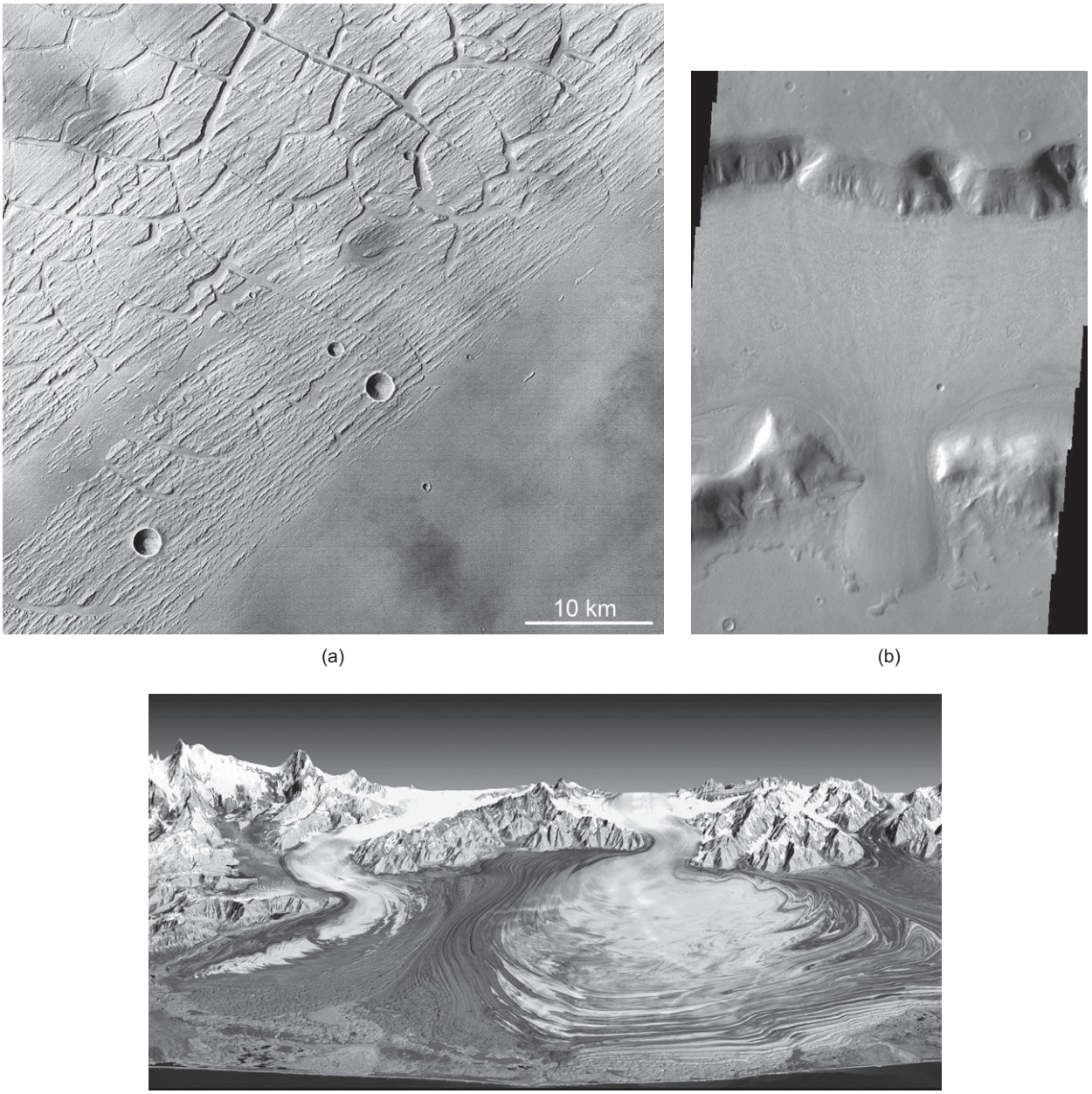


Figure 15.6 Examples of unusual Martian features interpreted to be glacial in origin. (a) Scour marks in Kansei Vallis, appear to be due to glacial erosion rather than by water erosion. The youthful nature of this area suggests that the glacial activity may have been recent. *Mars Express* image from ESA/DLR/FU Berlin (G. Neukum). (b) Piedmont lobe, about 3 km across, seen in Northern Arabia Terra. Such lobes are glaciers flowing out of a confined valley into a broad plain. Image from the Themis instrument aboard *Mars Odyssey*, from NASA/ASU (P. Christensen). (c) A terrestrial equivalent, the Malaspina Glacier in Alaska, is actually the merger of several glaciers. Landsat thematic image, courtesy SRTM Team NASA/JPL/NIMA. See color version in plates section.

15.3 Was Mars really warm in the past?

15.3.1 Limits to a carbon dioxide greenhouse

The picture of a warm early Mars is drawn by analogy with the early Earth – a thick carbon dioxide atmosphere sustaining a greenhouse effect in the face of a faint early Sun. Because of Mars' greater distance from the Sun compared to the Earth's – yielding only half the sunshine that Earth receives – a higher carbon dioxide pressure is required to sustain a certain temperature at any given epoch in the Sun's history. At least several atmospheres worth of carbon dioxide, or more, were required for Martian surface temperatures to be above the freezing point of water early in its history.

As shown in Figure 15.7, a potentially serious flaw arises for such a CO_2-only Martian greenhouse. For progressively smaller

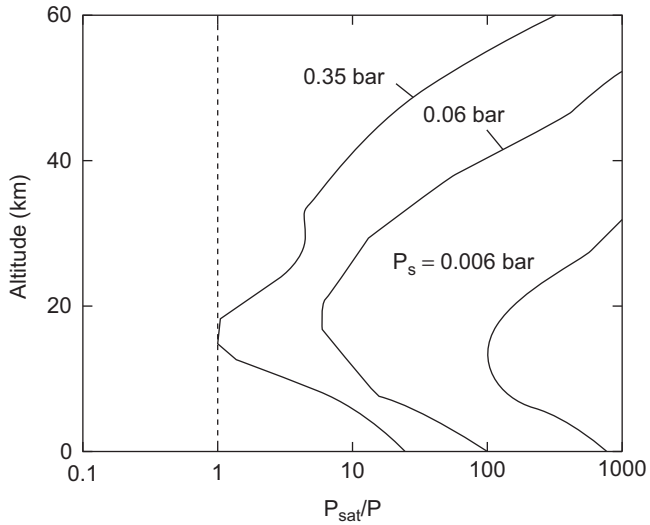


Figure 15.7 Greenhouse problem for Mars. The amount of carbon dioxide in the Martian atmosphere is shown for various values of the carbon dioxide surface pressure. Each profile is the ratio of the saturation pressure of carbon dioxide to the actual pressure as a function of altitude; where this ratio equals one (vertical dashed line), cloud formation occurs. The present-day carbon dioxide pressure near the surface (0.006 bar) leads to an atmosphere that does not produce carbon dioxide clouds (except near the poles). To get a Martian surface warm enough for liquid water, over 2 bars of pressure is needed at the surface (5 bars during the faint-Sun epoch). However, for any surface pressure above 0.35 bar, according to the figure, cloud formation will occur. Reproduced from Kasting (1991) by permission of Academic Press.

amounts of sunlight, one requires a higher carbon dioxide pressure to sustain a given atmospheric temperature. Carbon dioxide, like water, can form clouds, though much lower temperatures are required for a given amount of carbon dioxide to condense than for the same amount of water. It is possible to plot the carbon dioxide pressure for various temperatures at which carbon dioxide cloud formation will occur – the lower the temperature, the lower the pressure at which such clouds will form. An equivalent curve for water determines the altitude at which water clouds form in Earth's atmosphere for particular conditions on any given day. Figure 15.7 shows that CO_2 cloud formation occurs on Mars, for present solar luminosities, when the surface pressure of carbon dioxide exceeds 0.35 bar. This is many times less than the pressure of carbon dioxide needed to warm the surface to the water melting point, and is a direct consequence of Mars' greater distance from the Sun compared with Earth's. The carbon dioxide pressure needed to keep liquid water stable on *early* Mars is higher still (since the Sun was fainter), ensuring that CO_2 cloud formation must be considered in models of a warm early atmosphere.

The effect of carbon dioxide clouds on the early Martian greenhouse is not fully understood, because both scattering of sunlight and absorption of infrared energy (heat) may occur within such clouds. Most simple models suggest that the net effect of the clouds is to cool the atmosphere. This, in turn, requires higher carbon dioxide pressures to achieve a given surface temperature, which cannot be obtained because the gas simply condenses into thicker and thicker clouds, and eventually to carbon dioxide snow or rain. More elaborate models allow for both warming and cooling effects of clouds, so that cloud formation no longer is a hard limit to a CO_2 greenhouse effect. Francois Forget of Université Pie et Marie Curie, Paris and Ray Pierrehumbert of the University of Chicago found that clouds made of large particles of CO_2 ice can actually warm the Martian surface. Predicting particle size in clouds is difficult, but it is at least possible that a greenhouse atmosphere on early Mars was supported by CO_2 clouds with this property. Carbon dioxide pressures up to 2 bars can be contemplated, at which point the gas in the atmosphere may begin to reflect so much sunlight back into space that the greenhouse heating becomes inefficient.

One plausible way to circumvent the problem of warming early Mars is to posit other greenhouse gases that enhance the effect of the carbon dioxide. Water vapor is not a candidate, because the low temperatures at which the carbon dioxide cloud formation occurs are such as to keep the atmosphere extremely dry (water condenses out at much lower pressure, for a given temperature, than does carbon dioxide). What is required is a gas that condenses out at much higher pressure, and is a good infrared absorber. Methane (CH_4), ammonia (NH_3), and various compounds of sulfur – particularly sulfur dioxide (SO_2) have been proposed. The first two could plausibly be present only in the very earliest history of Mars, perhaps the first few million years, because they would quickly be broken apart by sunlight or surface reactions to form other species. The sulfur compounds might be more stable, and would be consistent with the large amount of sulfate deposit seen on the Martian surface today.

An alternative view is that the Martian surface was never stably warm for long periods of time, and that frequent impacts of asteroidal and cometary fragments released water and carbon dioxide into the atmosphere, allowing rainfall and formation of valley networks. A recent reanalysis by Brian Toon and colleagues of the timing of the formation of impact basins – large craters – on Mars, relative to the valley networks, suggests that they could have been causally related, and that impact-generated atmospheres thick enough to produce a greenhouse warming – while geologically brief in duration – would have allowed multiple flooding events occurring over centuries. The total number of such impact-driven clement episodes might have been sufficient to carve the channels. Only impacts early in Mars' history would have been effective in raising greenhouse atmospheres, because eventually the CO_2 required would have been depleted by loss to space and formation of crustal carbonates. Whether such a model is consistent with the evidence for long periods of standing water suggested by some of the Mars rover results remains an open issue, in part because the environment within which the liquid water was present remains poorly understood.

15.3.2 Abodes for life on Mars

Where might life have begun on early Mars? In the presence of a dense atmosphere, running or standing water at the surface might have provided environments suitable to the formation and maintenance of life. However, if standing or running water were transient or episodic, conditions obtained for a declining or marginally thin atmosphere, the environment would have been too unstable for prebiological chemical reactions to build in complexity and generate sustained, biochemical systems. The

presence of phyllosilicates, as noted earlier, does suggest an extended period of standing liquid water on or within the surface of Mars, enough for life to have formed based on when the earliest chemical traces of life appeared on Earth.

A stable liquid water environment need not mean that the liquid reached all the way to the Martian surface. NASA Ames scientist Chris McKay and others have studied lakes in Earth's Antarctic continent that are covered by a layer of ice year round. They sustain photosynthesizing algae. The liquid region below the ice is maintained in large measure by the warming effects of sunlight, transmitted through the clear ice into the liquid water below. In this respect, the ice-covered lake is analogous to a greenhouse atmosphere. However, the lake liquid also is stabilized by the warming effect of freezing itself, which releases heat. (This is the converse process to the cooling of a drink by the melting of ice cubes.) Calculations and field observations in the Antarctic suggest that such lakes are stable for temperatures as low as 240 K, fully 33 K below the freezing point of water.

Another possible birthplace of life is hydrothermal systems in the early Martian crust, regions where liquid water circulates in the rock and is warmed both by heat flowing from the interior and by the insulating effects of being underground. At Earth's mid-ocean ridges, hydrothermal systems are rich in the chemical and thermal energy needed to support an array of living organisms in the complete absence of sunlight. Such could have been the case on early Mars.

It is difficult to estimate how long either of these two types of ecosystems – surface bodies of liquid and hydrothermal systems – might have lasted on Mars. The groundwater hydrothermal systems are particularly problematic in this regard; we simply do not understand the details of Martian geologic history sufficiently well to predict where such systems could have been most long lived.

With regard to the ice-covered lakes, McKay and colleagues argue that they could have been maintained for some 700 million years after mean annual surface temperatures fell below the freezing point on Mars. If, for argument's sake, Mars had a warm climate continuously or episodically up through 3.5 billion years ago, then ice-covered lakes might have contained liquid water as late as 2.8 billion years ago. At this time on Earth, life was still solely in the form of single-celled prokaryotes, and the dramatic changes in our atmospheric composition wrought by such organisms (Chapter 17) had yet to take place.

We thus expect that any life on Mars would have remained at the single-celled stage at the time of its extinction. Life might have been sustained to the present day if the life formed in surface lakes or hydrothermal systems transitioned to deep environments in the crust, where liquid water may still exist today. The recent discoveries of bacteria living several kilometers beneath the surface of the Earth, in rocks that allow access to water and nutrients, hint at possible environments for life on present-day Mars. Perhaps beneath the surface of Mars, warmed by deep interior, simple Martian biota carry on.

15.3.3 Searching for evidence of life, and the early climate

It is a daunting task to explore other worlds, and even more so to search for evidence of microscopic life-forms that might be rare or hidden in inaccessible places. The search for evidence of past Martian life would involve looking for physical evidence (for example, stromatolites in ancient sediments), chemical evidence (ratios of carbon or other isotopes in rocks that are unusual except in biological processes), and mineralogical evidence (types of minerals that are not normally found together, except when formed through the mediation of biological activity). The search for extant life requires techniques to stimulate and detect metabolic activity or (more difficult) to isolate, reproduce, and then study the genetic materials of living organisms. The more exotic a Martian organism is to Earth life, the more challenging is its detection. Such searches must be conducted in regions of Mars where life was most likely to have formed and been sustained, and these are invariably the most difficult and dangerous sites to reach. For cost reasons, the search will continue to be conducted using robotic vehicles, most likely rovers as has been done with *Spirit* and *Opportunity*, but also with sensitive atmospheric sensors on orbiters – an approach that has already led to an intriguing result in the discovery of CH_4 (methane) in the Martian atmosphere.

Observed both by the European *Mars Express* orbiter and large Earth-based telescopic observations, methane was released into the Martian atmosphere in 2003 at levels of about 20 to 30 parts per billion. The methane seemed to be clustered over three different areas, including Nili Fossae, which is found to contain phyllosilicates (water-bearing minerals). By 2006 the abundance of methane had declined by a factor of two or more, and it has not been observed since. Although methane is destroyed by ultraviolet light in the Martian atmosphere, this process takes hundreds of years, and the rapid decline of the methane is a mystery (some have even suggested that the original observations were mistaken). If methane really was present in the Martian atmosphere, it could be the signature of biological processes, plausibly deep in the Martian crust where water and minerals might be present to sustain metabolisms. However, methane can also be produced by the reaction of carbon dioxide with water, in the presence of certain types of minerals, in a process called "serpentinization". Methane produced in this way could be a food source for life in the Martian crust, but would not be evidence of such life. A possible discriminant between biologically and nonbiologically produced methane is the ratio of two stable isotopes of carbon $^{13}C/^{12}C$, which differ in organic molecules processed by living versus nonliving systems (see Chapter 6), but the measurements from *Mars Express* and Earth-based telescopes are not sensitive enough to measure the isotopes. A future mission to Mars will be required to do this.

Although the discovery of evidence for past life on Mars would be a profoundly moving moment in human history, we must be cautious in how such a discovery would be interpreted. The existence on Earth of rocks blasted off of Mars in impact events shows that Earth and Mars have exchanged material over their histories. If life first formed on Mars, it might have seeded the Earth with life by this process of "impact exchange", or (less likely because of Earth's higher gravity) vice versa. To determine whether life on Earth and Mars had separate origins – a truly momentous discovery – would likely require analyzing preserved DNA from Martian organisms in terrestrial laboratories. Faint chemical signatures of past life on Mars could not establish whether or not it had a separate origin from life on Earth (Figure 15.8).

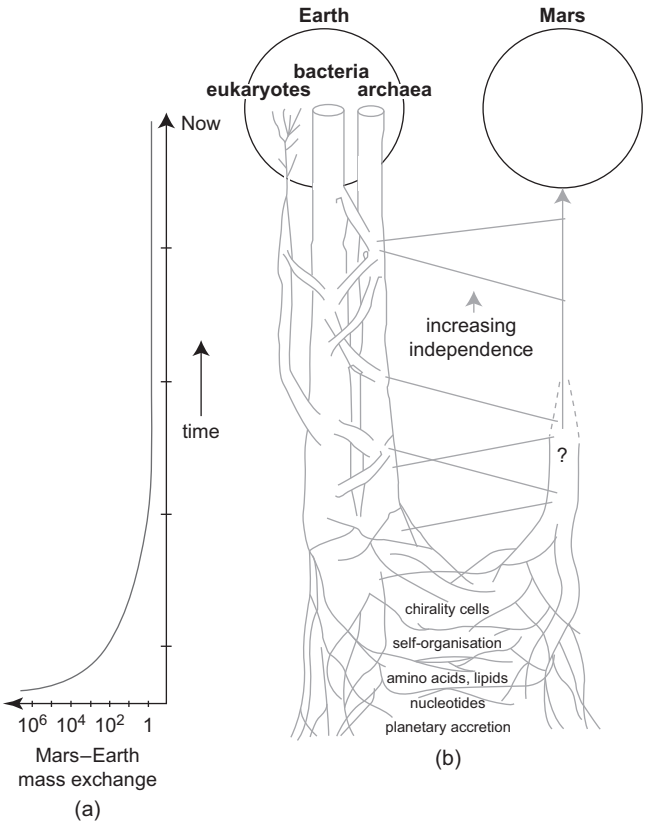


Figure 15.8 Conceptual exchange of genes between life on Earth and putative life on Mars by hypervelocity impact. Because the rate of impacts was much larger early in the history of these planets, the rate at which one planet's biota was contaminated by the other was larger than it is today. The left hand graph shows the rate of impacts with time normalized to a value of 1 for today. The right hand graph shows in a cartoon form the result: since a small amount of material deep inside a meteorite may not be shocked or heated to lethal values, organisms transported off one surface may survive landing on the destination planet. Thus Mars and the Earth have contaminated each other numerous times over their history, more frequently near the beginning. Life on Earth, and life on Mars if it exists, may have had a common origin. Figure provided by C. H. Lineweaver, from Lineweaver (2004).

Whether we find life on Mars – extinct or extant – or fail to do so, exploration of Mars is important for other reasons. A vigorous campaign to do so will pay off because the nature and demise of putative warm conditions on early Mars remains an unsolved problem. To solve it would be to gain a deep insight into the remarkable stability of our own planet's climate over four billion years.

15.4 Putting a Martian history together

Although speculative, here is a possible interpretation of the spacecraft data from Mars: after formation of Mars and an early epoch of heavy cratering, a period began that was characterized by high carbon dioxide abundance, the possible presence of other greenhouse gases, warm temperatures, and liquid water. This may have lasted for half a billion years or more, and primitive life might have formed. Over time, the CO_2 in the atmosphere was progressively locked up in surface carbonates or lost by impacts, and atmospheric pressure and temperature decreased. As temperatures dropped below the freezing point, glacial erosion became more important than erosion by liquid water. Perhaps several subsequent warm epochs occurred as chance large impacts or volcanic activity broke the crust, releasing water and carbon dioxide back into the atmosphere and allowing liquid water on the surface. Eventually, however, these reprieves ended. The atmospheric carbon dioxide continued to be progressively locked up in the sediments, and water ice became trapped in polar caps, as permafrost, and in the deep crust, in a process that continues up to the present cold, dry state. Life, initially retaining a toehold in lakes capped by water ice, or in subsurface hydrothermal vents, eventually ran out of sustainable climates and became extinct or relegated to a few sites deep below the surface – well hidden from the prying eyes and other sensors of Earth's robotic emissaries (Figure 15.9).

Figure 15.9 The Mars of today, captured by the Mars Exploration Rover *Opportunity* as it approached the rim of Endeavor crater in October 2011. The dry and uninhabitable landscape shows outcrops of rocks that contain evidence for an ancient epoch of warmer and wetter conditions. Image courtesy of NASA/JPL-Caltech.

15.5 Implications of Venusian and Martian history for life elsewhere

The search for evidence of life beyond our solar system is among the most daunting technological challenges imaginable. Direct evidence of life could come in the form of a radio beacon or even a visitation, but only from living forms advanced enough to do so and motivated enough to make contact. Absence of evidence of such contact is not evidence of absence of life-forms, by any means. An alternative approach, to examine neighboring

systems for planets of the right size and in the right location for harboring life, will yield interesting results even in the case in which few or no such planets are found (the conclusion then being that Earth is a rare pearl).

The exploration of Mars and Venus, and their scientific results, provide a framework within which to estimate the regions around neighboring stars wherein habitable planets might be found. Habitable is defined by most planetary scientists as being capable of harboring liquid water. Because evidence exists for liquid water on Mars in the distant past and isotopic data suggest an ocean on primordial Venus, one might argue that the range of habitability around stars like the Sun is from 0.7 to 1.5 AU. However, life must have time to develop and evolve; if we are interested in advanced life-forms we must look for planets with stable climates for liquid water over billions of years. Because the Sun is a typical main sequence dwarf, most main-sequence stars should similarly increase their luminosity over time, and hence the zone of "continuous" habitability around each star must be much narrower than the current Venus–Mars range.

We consider for the moment a star like the Sun, that is, of similar mass, composition, and luminosity history. The requirement of abundant liquid water early in a planet's history dictates that the outer edge of the habitable zone be inward of the current Martian orbit because, during the faint youth of the star, any planet at Mars' distance may have the same difficulty in sustaining a greenhouse as Mars had, regardless of the planet's size. Computations by Pennsylvania State scientist Kasting suggest that an early carbon dioxide greenhouse is readily sustainable within 1.15 AU of a Sun-like star; this could be extended outward for planets with other greenhouse gases or "warming" CO_2 clouds.

The inner edge of the habitable zone must be much closer to Earth than it is to Venus, because Venus suffered a moist greenhouse loss of water early in its history, and as a solar-type star heats up, planets progressively more distant than 0.7 AU will suffer the same fate. Computations by Kasting and others suggest that planets inward of 0.95 AU will suffer a moist greenhouse crisis for a luminosity of the central star equal to that of the Sun today. Hence, for a planet to sustain a liquid water surface over 4.5 billion years – the current age of the solar system and the length of time for sentient life to develop on Earth – it must be beyond about 0.95 AU from its Sun-like star.

The resulting zone of continuous habitability extends from 0.95 AU to 1.15 AU, that is, a width of 0.2 AU. What is the likelihood of finding a planet in another solar system orbiting at that distance from its parent star? In our own solar system, four rocky planets orbit the Sun between 0.4 and 1.5 AU. The mean spacing of the planets – four planets over 1.5 AU – is about 0.4 AU. If our system is typical, we can say that the likelihood of other planetary systems having a continuously habitable planet is just the width of the required zone divided by the typical (defined by our system) mean spacing, or $0.2/0.4 = 0.5$. This is a probability of 50% – quite high indeed!

Of course, other factors must come into play. Other systems have planets whose sizes differ from those in our solar system. Many stars are known to have planets comparable to the mass of Jupiter within the terrestrial planet zone (as defined by our solar system). A system in which a Mars-sized body occupied Earth's orbital position might not have hosted life because such a small planet would not possess the plate tectonic recycling of carbon dioxide needed to sustain an environment stable for liquid water. Conversely, a more massive planet possessing plate tectonics but occupying the orbit of Mars might develop a sustainable greenhouse atmosphere, possibly a bit later than did our Earth as its parent sun increased in luminosity, and then could sustain a clement climate over billions of years.

The mass of the central star itself also determines habitability. More massive stars are more luminous and hence habitable planets must be in proportionately larger orbits than Earth's orbit around the Sun. However, more important, the higher luminosity comes from a more rapid rate of hydrogen fusion, so that massive stars are shorter lived (Chapter 4) and hence provide their planets a much smaller time for biological evolution. Happily, the most abundant stars in the galaxy are the least luminous and longest lived: the M-dwarfs. They, however, harbor another difficulty: for a planet to gain sufficient warmth from an M-dwarf to be habitable requires that it be as close to its star as Mercury is from the Sun. In that tight orbit, gravitational tugging will force the planet toward a state of *synchronous rotation* where one side faces toward the star – just as our Moon keeps one face toward the Earth. The resulting climate will be vastly different from Earth's, and perhaps not conducive to stable liquid water. Flares and "coronal mass ejection" of material from the star's atmosphere will erode the planet's atmosphere, unless is possesses a powerful magnetic field to deflect charged particles. In general, the habitability of a world tucked in close to its parent star, as is the case for M-dwarfs, is questionable and remains poorly understood.

Returning to our own solar system, human technology might expand the habitable zone in the not-too-distant future. A number of futurists have proposed seeding the Martian atmosphere with efficient greenhouse gases such as methane and chlorofluorocarbons, in an effort to warm the surface, release water, and generate conditions that are more Earth like. Although beyond the reach of current space transportation capability, such a *terraforming* of Mars might be possible for a future generation, which will have to weigh the advantages against the potential ethical dilemmas associated with transforming a vast natural environment.

15.6 The finite life of our biosphere

The evolution of our Sun has one more consequence for life on Earth. From now to the end of its stable hydrogen-fusing stage, the Sun will continue to increase in luminosity. As it does so, the climate of the Earth will edge closer to the point at which a moist greenhouse is initiated and rapid loss of the Earth's water ensues, as apparently occurred early in Venus' history. A natural delaying tactic is the weathering feedback process described in Chapter 14 wherein, as the brighter Sun warms Earth, more rainfall and more erosion will occur, and hence the carbon dioxide budget of our atmosphere will decrease. However, a point will come when rising temperatures cannot be buffered by the decreasing amounts of atmospheric carbon dioxide, and rapid loss of Earth's oceans to space will begin.

Models of the Sun's luminosity history and the response of Earth's atmosphere suggest that this crisis will be reached in

1 billion to 2 billion years from now. At that point, if the biosphere has not collapsed already from decreasing amounts of atmospheric carbon dioxide, the lack of liquid water will finally kill off all living organisms. On the other hand Mars will enjoy more clement conditions, if enough water and carbon dioxide are stored in the crust to be partially liberated into a thicker atmosphere by the brighter sun.

Life began on Earth some 3.8 billion to 4 billion years ago, and complex eukaryotic cells appear in the fossil record from 2 billion years ago. Therefore, we are more than halfway through the time period during which life, even complex life, can flourish on Earth. Our time here is not forever. After the brightening Sun drives water, and hence life, from Earth, it will continue to shine by hydrogen fusion for another 2 billion to 4 billion years. For those last several billion years of the Sun's history, Earth's surface might hold a fossil record of its long springtime of clement conditions, during which it teemed with living organisms that eventually looked upward to contemplate the stars.

Summary

Earth's two neighboring planets are very different from each other in size and climate history. Venus, not too different from the Earth in size, orbits 30% closer to the Sun than does the Earth. One would expect Venus to be hotter than Earth, but it is the enormous atmosphere of carbon dioxide and consequently massive greenhouse effect that is most striking about this planet. Indeed, though less sunlight arrives at the surface of Venus than arrives on our home world's surface, thanks to Venus' global layer of clouds, the temperature there is 730 K – hot enough to melt lead. Isotopic evidence suggests that Venus was once clement enough to support large amounts of liquid water at or near its surface, water that was lost in an environmental crisis precipitated by our sister planet's proximity to the Sun. When that happened is not known, but it left Venus with no water, a surface too hot to support liquid water and life, and a very different geologic style than the Earth's plate tectonics. Venus is an example of what happens to an Earth-like body at the inner edge of the "habitable zone". Mars, on the other hand, is farther from the Sun than is the Earth, and is ten times smaller. It has a very thick carbon dioxide atmosphere with essentially no greenhouse effect, and a surface that is bombarded by ultraviolet light and is on average much too cold to support liquid water. But there is ample evidence from images, spectra, and chemical data, and from orbiters and rovers on the surface, that liquid water existed on the Martian surface and just below the surface in the ancient past. How Mars could have sustained clement conditions early on when the Sun was significantly fainter than today, and for how long such conditions existed, remain unsolved problems. If Mars was clement early in its history, life might have begun on the surface and then found refuge in the deep crust, where liquid water is still stable today. The detection of small amounts of methane in Mars' atmosphere for a period of several years suggests either biological activity or the action of water on carbon dioxide in the deep crust, but the transient nature of the detection makes difficult its eventual follow up by future missions to determine the source. As Venus and Mars define in a coarse way the outer edges of the habitable zone, the continuing evolution of the Sun toward higher and higher luminosities means that, in 1 to 2 billion years, Earth may lose its water and become uninhabitable. Studying Mars and Venus to understand the limits of habitability will help us to understand how precarious is the habitability of our own planet.

Questions

1. How and where would you search for evidence of past, and present, life on Mars?
2. Can you think of any refugia for carbon-based life on present-day Venus?
3. Speculate on what the evolution of an Earth-mass planet might be if it were placed at the orbit of Mars around the Sun. Would it be better able to retain greenhouse gases than did Mars? What would be the tradeoff between more mass allowing plate tectonics to occur, against the much weaker amount of sunlight in a Mars-like orbit? Might such a planet start out frozen, become habitable, and avoid the runaway crisis the Earth will face?
4. Sulfur compounds have been suggested as a possible greenhouse gas for early Mars. What sulfur molecules might be suitable? Would they condense? Could they form a smog that might impede sunlight? What is the difficulty with methane as an additional greenhouse gas?

General reading

Nimmo, F. and McKenzie, D. 1998. Volcanism and tectonics on Venus. *Annual Reviews of Earth and Planetary Sciences* **26** DOI: 10.1146.

Pierrehumbert, R. T. 2010. *Principles of Planetary Climate*. Cambridge University Press, Cambridge.

Squyres, S. W. 2006. *Roving Mars: Spirit, Opportunity and the Exploration of the Red Planet*. Hyperion Press, New York.

References

Ehlmann, B. L. 2010. Diverse aqueous environments during Mars' first billion years: the emerging view from orbital visible-near infrared spectroscopy. *Geochemical News*, 142.

Caldeira, K. and Kasting, J. F. 1992. The life span of the biosphere revisited. *Nature* **360**, 721–3.

Forget, F. and Pierre Humbert, R. T. 1997. Warming early Mars with carbon dioxide clouds that scatter infrared radiation. *Science* **278**, 1273–6.

Holt, J. W., Safeinili, A., Plaut, J. J. *et al.* 2008. Radar sounding evidence for buried glaciers in the southern mid-latitudes of Mars. *Science* **322**, 1235–8.

Kargel, J. S., Baker, V. R., Begé, J. E. *et al.* 1995. Evidence of ancient continental glaciation in the Martian northern plains. *Journal of Geophysical Research* **100**, 5351–68.

Kasting, J. F. 1988. Runaway and moist greenhouse atmospheres and the evolution of Earth and Venus. *Icarus* **74**, 472–94.

Kasting, J. F. 1991. CO_2 condensation and the climate of early Mars. *Icarus* **94**, 1–13.

Kasting, J. F. 1997. Planetary atmosphere evolution: do other habitable planets exist and can we detect them? In *The Search for Extra-solar Terrestrial Planets. Techniques and Technology* (M. Shull, H. Thronsoan and A. Stern, eds). Kluwer, Dordrecht, pp. 3–24.

Kasting, J. F., Whitmire, D. P., and Reynolds, R. T. 1993. Habitable zones around main sequence stars. *Icarus* **101**, 108–28.

Lineweaver, C. H. 2004. Martian life: stuck somewhere between inevitable biochemistry and quirky biology. *Microbiology Australia* **25**(1) 20–1.

McKay, C. P. and David, W. L. 1991. Duration of liquid water habitats on early Mars. *Icarus* **90**, 214–21.

Mumma, M. J. Villanueva, G. L., Novak, R. E. *et al.* 2009. Strong release of methane on Mars in northern summer 2003. *Science* **323**, 1041–5.

Phillips, R. J. and Hansen, V. L. 1994. Tectonic and magmatic evolution of Venus. *Annual Review of Earth and Planetary Sciences* **22**, 597–654.

Rampino, M. R. and Caldeira, K. 1994. The Goldilocks problem: climatic evolution and long-term habitability of terrestrial planets. *Annual Reviews of Earth and Planetary Sciences* **32**, 83–114.

Sackmann, I-J., Boothroyd, A. I., and Kramer, K. E. 1993. Our Sun. III Present and future. *Astrophysics Journal* **418**, 457–68.

Sagan, C. and Mullen, G. 1972. Earth and Mars: evolution of atmospheres and surface temperatures. *Science* **177**, 52–6.

Squyres, S. W. and Knoll, A. H. 2005. Sedimentary rocks at Meridiani Planum: origin, diagenesis, and implications for life on Mars. *Earth and Planetary Science Letters* **240**, 1–10.

Toon, O. B., Segura, T., and Zahnle, K. 2010. The formation of Martian river valleys by impacts. *Annual Review of Earth and Planetary Science* **38**, 303–22.

van Thienen, P., Vlaar, N. J., and van den Berg, A. P. 2004. Plate tectonics on the terrestrial planets. *Physics of the Earth and Planetary Interiors* **142**, 61–74.

Zuber, M. T. 2001. The crust and mantle of Mars. *Nature* **412**, 220–7.

16

Earth in transition: from the Archean to the Proterozoic

Introduction

The beginning of the Proterozoic eon is set formally by geologists at 2.5 billion years before present. However, the transition between the Archean and the Proterozoic is not a sharp one. From about 3.2 billion to 2.5 billion years ago, rocks with a modern granitic composition made a widespread appearance in the geologic record. Prior to this time, rocks making up the Archean continents had a composition different from modern granites in several important respects. Beginning around 3.2 billion years ago in what is now Africa, and extending to 2.6 billion years ago on the Canadian shield, large quantities of modern-type granites were produced. We can collect these rocks today and date them by use of radioisotopes. How did the original Archean continents form? Why was there a transition in chemical composition of the rocks roughly halfway through the Archean? What might Earth have been like today if this eruption of new rock types had not occurred? As we see, the transformation wrought on Earth's primitive continents may have been an inevitable consequence of their increasing coverage of Earth's surface.

What might have been inevitable on Earth was apparently difficult or impossible on the other terrestrial planets. No evidence for large granitic masses exists on any other planet. Venus bears two crustal masses that resemble continents, but the details of their geology suggest that they are more similar to primitive Archean continents than to our modern ones and, even then, the connection is a weak one. Venusian geology, taking place as it did on a planet similar in size and composition to Earth, might teach us about the conditions under which Earth's tectonic regime could *not* be achieved.

The formation of continental masses, standing above the level of the seas on an otherwise watery world, had a profound influence on the subsequent history of Earth. The weathering of continental granites provides the essential first step in the transport of atmospheric carbon dioxide into the oceans and sequestration as carbonates. The continents as buoyant regions of the crust probably largely determined the pattern of plate tectonics. They modulate the climate and interior heat flow through cycles of merging into single supercontinents and breaking up into dispersed land areas. Unlike the oceanic crust, which is destroyed and recreated on timescales of a few percent of Earth's history, the continents preserve an ancient record of the geologic and biological history of Earth. Finally, the continents provided a new frontier for colonization by complex life some half-billion years ago, an environment in which the nature of survival differed drastically from that in the sea.

It is in the Archean and its transition to the Proterozoic that we see Earth, once and for all, diverge in its evolution from that of its sister planets and become the planet with which we are familiar. Understanding how this happened is the subject of the present chapter.

16.1 Abundances of the elements in terrestrial rocks

Table 16.1 summarizes the amounts of the most abundant elements in typical basaltic and granitic rocks, compared to numbers for the mantle of Earth, and for the most primitive class of meteorites (carbonaceous chondrites). Granitic-type composition dominates continental crust, whereas basalt is typical of oceanic crust. Abundances of the major elements vary quite a bit from rock to rock, particularly since continental rock is by no means purely granitic. Hence the numbers in the table are illustrative; what is important are the general trends. The mantle composition is the least certain, because it depends largely on volcanic rocks whose formation in the mantle and subsequent chemical alteration are not precisely known.

Table 16.1 Chemical composition of rocks[a]

Element	Chondritic meteorites	Earth's mantle	Basalt	Granite
O	32.3	43.5	44.5	46.9
Fe	28.8	6.5	9.6	2.9
Si	16.3	21.1	23.6	32.2
Mg	12.3	22.5	2.5	0.7
Al	1.4	1.9	7.9	7.7
Ca	1.3	2.2	7.2	1.9
Na	0.6	0.5	1.9	2.9
K	0.1	0.02	0.1	3.2
Other	5.9	1.7	2.7	1.6

[a] Composition in percent by weight, e.g., 33% oxygen means one-third of the weight of the rock is oxygen. Data from Broecker (1985).

Some interesting systematics arise from the table, beyond the dominance of oxygen, the most abundant element in the rocks. Iron is very abundant in the chondrites but much less so in the three types of Earth rocks. This is understood to be the consequence of the formation of a largely iron core very early in Earth's history (Chapter 11). Silicon is commensurately more abundant as iron declines. Magnesium, which is very abundant in mantle rock, declines drastically in both basalts and granites. Other elements are more abundant in the crustal rocks than in the mantle or the chondrites. These include aluminum, sodium, and potassium. Finally, between the crustal rocks, some further differences emerge: basalts contain more iron, magnesium, and calcium than do the granites, whereas sodium and potassium are strongly enhanced in the latter.

We can regard chondrites as roughly representative of the original chemical mix from which Earth formed – but only roughly, because there are differences in the details of the various chondrite classes both from each other and from the Earth. Comparing with chondrites, it is relatively straightforward to see how the mantle is a residuum of the removal of iron and elements that tend to follow iron. The pattern of the crust, however, seems less certain and, in particular, basalts and granites seem to have followed different paths in their formation. The discussion in Chapters 9 and 11 suggest that the crustal rocks ought to be derived in some way from the mantle, and the major elements should provide a guide to that process. To gain some insight into how this might have happened, we must discuss mineral structures and the concept of partial melting of rocks.

In addition to elements that are abundant in crustal rocks, certain trace elements are important indicators of the origin and history of continental rocks. Chief among these are the *rare-earth elements*, or rare earths, which do not dissolve very well in water. They tend, therefore, to stay with the rocks as the rocks are moved around in rivers as sediments. Furthermore, their abundances are an important distinguishing feature between Archean and post-Archean continental rocks. As can be seen by consulting the periodic table (Figure 2.6), the rare earths all occupy a single column in the table, but move upward in atomic number by filling an interior (fourth) series of electronic energy levels while the outermost (fifth) is complete (see Chapter 2). In consequence, these elements all share certain common chemical properties of diagnostic value in determining the origin and history of continental crust.

16.2 Mineral structure

The structure of minerals represents a scientific discipline in and of itself, and justice cannot be done in the short summary here. Like all chemical substances, the building blocks of minerals are the chemical elements, and the properties of minerals are determined by the types of elements present and how they are bonded together. The abundant elements in minerals are joined by ionic bonds, where electrons are actually transferred from a donator to an acceptor element. For example, in sodium chloride, NaCl, each sodium atom donates an electron to chlorine, and the two types of elements arrange themselves in a regular lattice structure that constitutes the crystalline structure observable on a macroscopic scale. Having donated an electron, the positively charged sodium (Na^+) is packed in between the negatively charged chlorine (Cl^-).

Most significant about ionic bonding is the donation of the electron, which produces ions of very different sizes. As noted in Chapter 2, the sizes of the elements across a given row of the periodic table vary slowly, determined by the presence of the electrons. Removal of an electron in ionic bonding greatly decreases the size of the resulting positive ion, whereas the accepting element, becoming a negative ion, increases in size. The *ionic radius* of an element refers to its size in a certain ionic state, that is, having donated or accepted an electron. The *atomic radius* (that is, of the neutral element) of sodium is almost twice that of chlorine, 1.86 Angstroms (1.86×10^{-10} m) versus 0.99 Angstroms. However, Na^+ has an ionic radius of 0.98 Angstroms, whereas Cl^- is 1.81 Angstroms. Thus, the size situation essentially is reversed in the ionic bonding to form sodium chloride.

The stability of a particular chemical compound, or mineral, lies in part in its structure, which in turn depends on the relative sizes (and hence ionic radii) of the elements. Mineral structures tend to follow a pattern where several negatively charged ions (*anions*) surround a positively charged ion (*cation*). Oxygen is a very common anion in minerals. In *calcite*, three oxygens surround one carbon (in addition to the presence of the element calcium); in olivine and other minerals, four anionic oxygens (O^-) surround one silicon (Si^{4+}) (in addition to two magnesium or iron cations, which also are present). The ionic radius of the silicon cation is roughly one-fourth that of an oxygen anion; hence, the *tetrahedral coordination* could be predicted essentially from size alone.

The enclosure of an element with a small atomic radius by several with large radii produces a well-packed, closed structure (Figure 16.1). However, not all cations and ions fit together. A large cation tends to push the anions outward, allowing a separation between them that is not the energetically most favorable situation. A cation that is too small allows the anions to just touch, but does not interact strongly enough with the anions to stabilize the configuration. Therefore, for any given number and type of anion, cations of particular size *and hence elements of particular ionic radii* optimally stabilize the crystal structure.

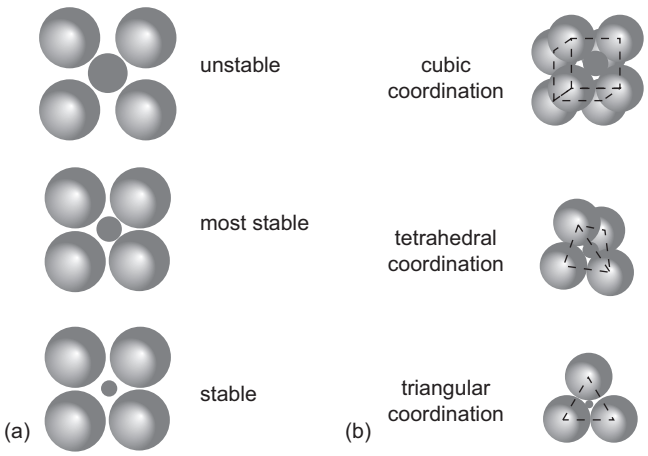


Figure 16.1 (a) Fit of small, medium, and large cations (dark gray) within mineral structures. (b) Some examples of typical arrangements or *coordinations* of anions in minerals. Adapted from Press and Siever (1978) and Broecker (1985).

Table 11.1 in Chapter 11 lists the ionic radii of the elements of Table 16.1. The sizes of elements in their cationic form determine their ability to be accommodated in major minerals. For example, aluminum as a cation is similar in size to silicon, and readily substitutes. Magnesium and iron are similar in their ionic radii, and both occur in silicate minerals such as olivine. However, calcium, potassium, and sodium have large ionic radii and cannot substitute for silicon in a four-oxygen structure; they require a different coordination with a larger number of oxygen atoms.

In the structural compatibilities and mismatches of various cations lies much of the reason for the chemical differences between mantle and crustal rocks, and between oceanic and continental crust. The process leading to such differences is the melting of rock beneath the surface of Earth.

16.3 Partial melting and the formation of basalts

The mantle of Earth is a solid with plastic properties, oozing slowly in a pattern of upwelling and downwelling motions that remove the internal heat of the planet. The solid nature of the mantle has been well established by seismic data and studies of mantle-composition rocks subjected to high pressures in the laboratory (Chapter 11). However, volcanism both on continents and at mid-ocean ridges amply demonstrates that at any given time some mantle material is melted, and this melt rises to form a part of the crust.

Clues to the process by which mantle melting occurs come from laboratory studies of common minerals, combined with some understanding of thermodynamics. On the surface of Earth, we are used to the melting of any given material being characterized by a single parameter, namely, temperature. Water ice melts at 273 K (32°F); common basalts melt at around 1,500 K. Melting, or passage into the liquid state from the solid, is a function also of composition of a mineral and, importantly, pressure. For most materials the liquid form is less dense, or more voluminous, than the solid; that is, the solid sinks in the liquid. Likewise, the melting temperature increases with pressure. (The exception is water, for which ice floats on the liquid, and the melting temperature decreases with pressure up to roughly 2,000 bars, equivalent to about 2 km depth in the ocean.)

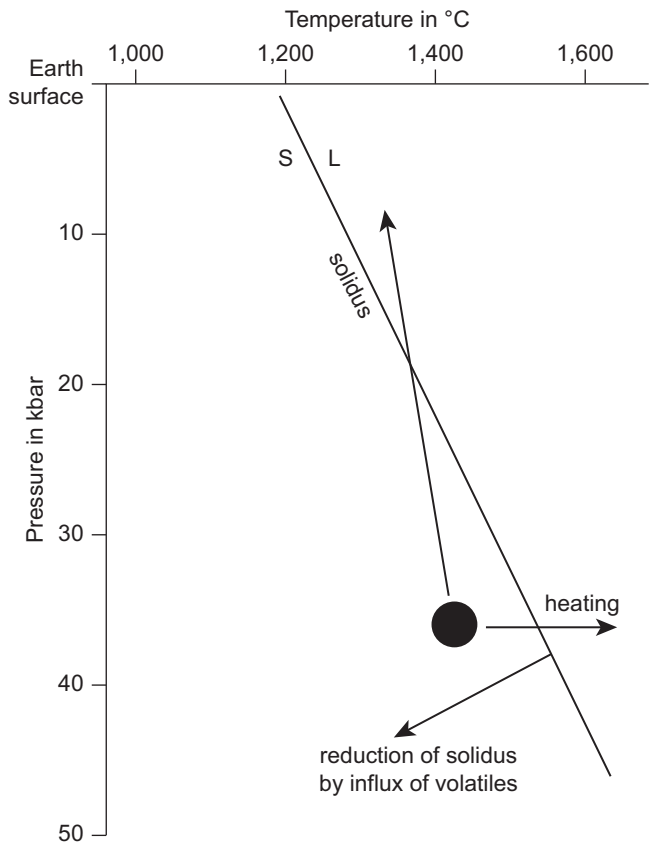


Figure 16.2 Schematic of how melting occurs in the mantle of Earth. The diagonal line is the boundary, or *solidus*, between having all solid (below and to the left) or solid with some liquid (partial melting); it is based on the chemical properties of the mantle rock. Imagine a piece of the mantle at a temperature and pressure given by the circle. Simply heating the sample will slide it across the solidus to the partial melt zone. However, more typical is that mantle material, moving upward, encounters decreasing pressure and temperature shown by the nearly vertical arrow. Although the temperature decreases along the curve, the sample crosses the solidus and melting takes place. The arrow pointing down and to the left shows the direction in which the solidus slides as more water is added to the mantle; melting is made easier. From Rogers (1993).

Imagine now being present in a part of the mantle of Earth that is rising toward the surface, carrying away interior heat. As this part of the mantle rises, temperature drops from high internal values toward the surface value. This brings the mantle material further away from its threshold of melting. However, the pressure also drops, and this tends to push the mantle material toward its melting point. The run of temperature with pressure in the mantle, based on computer models and heat-flow measurements at the surface, is such that the decrease of pressure is the dominant effect (Figure 16.2). That is, the rise of solid mantle material toward lower pressures allows some of the material to melt, in a process called *pressure-release partial melting*.

"Some" is the operative word here. The mantle rock consists of an assemblage of major and minor elements, located in various crystalline structures dominated by oxygen as anions and silicon, magnesium, and iron as cations. Melting does not occur wholesale in such an amalgam; rather, certain combinations of elements tend to preferentially move into the melt. This property of *partial melting* is not unique to melting rocks; water mixed with other materials (e.g., ammonia or organic solvents such as commercial antifreeze) undergo partial melting at a temperature well below 273 K.

What determines which elements move preferentially into the melt? One important determinant is discussed above: the size or ionic radius of the cations. Oversized cations (potassium, calcium, and sodium) that have trouble fitting into the four-coordinate silicon-oxygen structure are energetically favored to be preferentially in the melt. This is not an either/or proposition; some fraction of the larger cations will remain in the solid, retaining their eight or larger coordination with oxygen, but overall they tend to concentrate in the melt.

Silicon and oxygen are so abundant that they remain the dominant constituents in both the melt and the remaining solid. Magnesium tends to reside preferentially in minerals with higher melting points, and hence favors the solid phase as the melt progresses. Iron and aluminum have a slight preference for the melt. Reference to Table 11.1 in Chapter 11 shows that these cannot be size effects, but are related instead to other properties of the bonding mechanisms between the cationic elements and the anionic oxygen.

As melt forms, being just slightly less dense than the solid, it moves upward more rapidly toward the surface. In effect, the lower-density melt is buoyant in the surrounding solid (just as a hot-air balloon is buoyant in the surrounding cooler air). Thus, unlike a chemistry experiment in which a material is melted in the flask and solid and liquid phases continue to interact chemically, the mantle melt leaves its solid residue behind and moves upward through "fresh" mantle rock, some of which also melts. Pressure in the cracks through which the melt moves helps force it upward; this effect is particularly important at depths where the melt is just about the same density as the surrounding rock. By the time the melt reaches Earth's surface, it occupies a significant fraction of the total volume of the rock through which it moves.

Most of the mantle melt reaches the surface at mid-ocean ridges where it oozes out, forming new oceanic crust. The composition of the new material is basaltic, corresponding roughly to that given in Table 16.1. Sodium, potassium, iron, aluminum, and calcium are enhanced and magnesium is depleted in this melted derivative of the mantle. Not all basalts erupt at mid-ocean ridges. Some come to the surface at isolated *hot spots* around the globe, such as the Hawaiian islands. Much of this hot-spot basalt appears to have been melted at somewhat deeper levels than the mid-ocean-ridge basalts, with consequent interesting differences in element abundance.

The formation of basaltic crust represents a second stage of the chemical differentiation of Earth, the first stage being the primordial separation and sinking of iron to form a core, with loss of some fraction of additional elements that tend to combine with iron, such as sulfur and oxygen. The partial melting of rising mantle material as basalts is probably a process common to Earth, Mars, and Venus. What differs about Earth is that basalt is not the sole crustal material. Most of the continental crust is not basaltic in composition, and its composition is not derivable from the simple partial melting process considered here. Instead, the origin of continental rocks appears to be intimately tied up with the special feature of Earth as a planet, namely its abundance of liquid water.

16.4 Formation of andesites and granites

16.4.1 Rock relationships

To gain insight into the formation of terrestrial rocks other than basalts requires understanding the chemical relationships between the major rock classes. Figure 16.3 illustrates these relationships, for the major igneous rocks only; recall from Chapter 8 that metamorphic and sedimentary rocks begin their existence as igneous rocks. The primary distinguishing feature among the igneous rocks is their silica (or SiO_2) content. Felsic rocks (richer in SiO_2) are light colored, relatively low density; mafic rocks are poor in SiO_2, have proportionately more iron and magnesium, and are therefore higher density and darker color. There is a progression of rocks from felsic to mafic. This progression occurs along two tracks, corresponding to whether the rock erupted from a volcano, and hence is *extrusive*, or instead, the melt stopped rising within the relatively low-density continental crust (hence was no longer buoyant), and cooled to solidification as an *intrusive* rock.

The major extrusive/intrusive equivalent rock types, in progression of chemical composition from mafic to felsic, are basalt/gabbro, andesite/diorite, rhyolite/granite. The more common felsic rock class is the intrusive granites; the more common extrusive mafics are the basalts. In some of what follows, we rather loosely describe oceanic crust as basaltic, continental crust as granitic, and crust produced at continental margins above descending slabs as andesitic. Mantle rock is even more mafic than basalt; an example of this kind of rock is called *peridotite*.

16.4.2 Seismic waves and composition

As a brief aside, the seismic P-wave velocity (Chapter 11) changes smoothly with composition; granites have a P-wave velocity of 6 km/s, gabbro and basalts 7 km/s, and peridotite 8 km/s. The seismic-wave explorations described in Chapter 11 for Earth show that the oceanic crust is basaltic (along with some gabbro) with no granites, and the upper mantle is consistent with an ultramafic composition like that of peridotite. Interestingly, the thick continental crust is not entirely granitic; the P-wave velocities suggest that some gabbro is present in the lowermost continental crust. In this dual nature of continental crust lies further clues to continent formation.

16.4.3 Role of water in partial melting

As Table 16.1 shows, a granite is made from a basalt by reducing the iron and magnesium contents while increasing the sodium and potassium abundances. Andesites are, compositionally, an

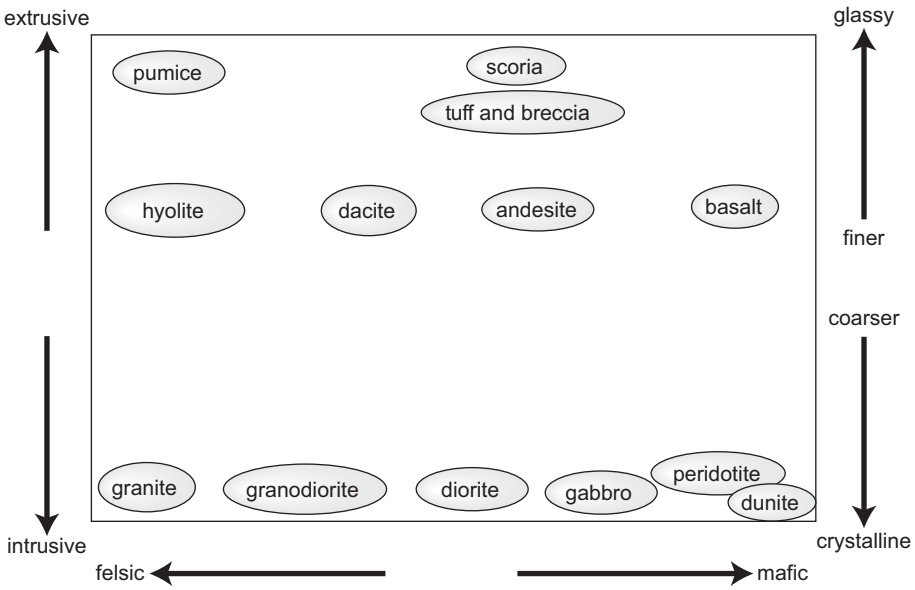


Figure 16.3 Relationship between selected major igneous rocks. Rock types are positioned according to the amount of silica they contain, and whether they are plutonic (intrusive) or volcanic (extrusive). Plutonic rocks do not breach the surface during eruptions, but cool and solidify buried in the crust. Because these rocks cool slowly, they are composed of large crystals. Volcanic rocks, which erupt to the surface, are cooled suddenly and have glassy textures, or very fine crystals.

intermediate step. The challenge lies in identifying the geologic environment in which large amounts of such chemical alterations can occur. Simply cycling the basalt back into the mantle is not sufficient to produce andesites and granites; the melting of dry basalt occurs at a fairly high temperature, thus fairly deep, in such a fashion that the liquid retains a composition quite similar to that of the original rock. (Recall that, because basalt is the solidified product of partial melting of mantle rock, it will melt under different conditions than the original mantle rock.)

What is required to make more felsic rocks is the addition of a material that will alter the melting relationship of the basalt, and water is an excellent candidate. Figure 16.2 shows that, in the presence of sufficient amounts of water, the melting point of the basalt drops dramatically. Further, the melting is partial, with larger ions again partitioning preferentially into the melt. The resulting liquid has an andesitic composition, partway between basalt and granite.

The physical environment in which water plays a role is supplied by plate tectonics (Figure 16.4). The basalt at the mid-ocean ridges solidifies as it approaches the surface. Because the ridges are underwater, the basalts react with the water, which becomes incorporated in the crystal structure of the rocks, *hydrating* the minerals. The process can be very efficient because the midocean ridges are filled with cracks, through which water circulates in an intricate network of hydrothermal systems. Thus, much like the cells in our body receive nourishment through an intricate network of capillaries, oceanic basalts enjoy extensive and intimate contact with water.

As oceanic crust moves away from mid-ocean ridges, it cools and thus becomes denser than the rock beneath. Sitting on less dense rock is an unstable situation, and the crust eventually founders and sinks into the mantle below. This sinking, or subduction, can take place at the boundary of a continent or purely within an ocean basin. Basalt that was not fully hydrated at the ridges now has a second chance as it is warmed up during its descent below the continental edge. However, temperatures continue to climb as the subducting *slab* descends, and eventually become too high for water to remain stably bound in the basaltic rock. The water becomes unbound from the minerals and, being buoyant, rises upward through the boundaries between mineral grains.

The release of water has two profound effects on the subducting slab and the adjoining mantle. First, the slab rock becomes denser, aiding subduction. Second, the water rises above the slab into the wedge of mantle above; as it contacts basaltic and mantle grains, the melting point of the grains plummets, and partial melting of the slab and mantle begins at temperatures and pressures much lower than that which originally formed the basalts. The partial-melt products are less dense than the mantle and hence rise, though more slowly than the water because the molten rock is closer in density to the mantle itself. The partial melt, andesitic in composition, erupts onto the surface in the form of volcanic lavas; some may come to rest near the surface as igneous intrusions of diorites.

Andesitic volcanism, or *arc volcanism*, is common along the margins of active subduction zones, for example, along the Andes, Japan, and the northwestern United States. Some of these margins are actually disconnected from the adjoining continent (as with Japan). Aside from an enrichment in sodium and potassium, the andesitic or dioritic rock is enriched in the large ions uranium and thorium. Because these elements, along with potassium, have relatively abundant, long-lived radioactive isotopes, the effect of the low-temperature melting of the mantle at plate margins is to concentrate the heat-producing elements in the andesitic crust.

The reader may notice an interesting correspondence between the cycling of water into the mantle and that of carbon, described in Chapter 14. Indeed, in both cases, plate tectonic subduction

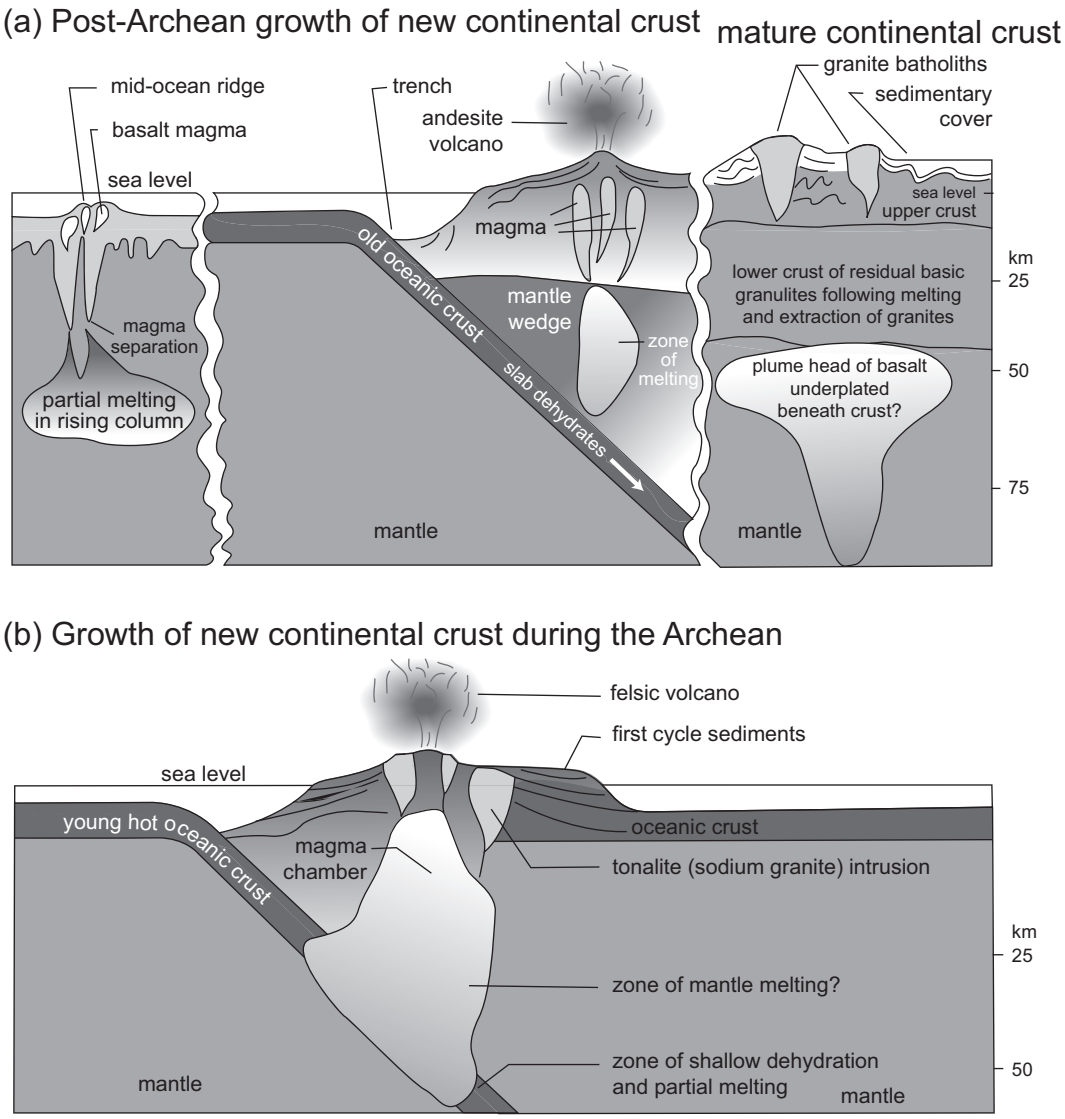


Figure 16.4 View of the formation of continental crust: (a) today and (b) in the Archean. The size of the continent is truncated in the top panel; only the edge and a portion of the interior continental shield are shown. At the continental margin, formation of andesite is fairly well understood. In the interior, it is suggested that the formation of granites results from melting of more mafic rocks in the lower continental crust, with the granites rising to the top as batholiths. In the bottom panel, it is speculated that Archean plate tectonics involve rapid recycling of ocean crust and small continental minishields, which are the products of partial melting of subducting slabs – made possible by the higher temperatures of the Archean mantle. However, because hydration of the basalt (that is, chemical combination with water) is not complete, and Archean temperatures were quite high even at shallow depths, the melt product is not andesite but unusual granite-like "granitoid" rocks rich in sodium, including the so-called (*tonalites*). From Taylor and McLennan (1995).

provides a mechanism for recycling volatiles trapped chemically in oceanic sediments (carbonates) or basaltic crust (water) back onto the surface. In the case of carbon, the recycling is key to sustaining a warming atmospheric greenhouse; in the case of water, its release from rock provides the key step in low-temperature partial melting of basalts and mantle to form andesites and diorites.

16.4.4 The puzzle of granite formation

The upper crust of the continents is not made predominantly of andesites or diorites. The granite of which it is composed is even further away in composition from basalt than are the andesites, and hence must represent an additional cycle of differentiation. However, no obvious simple process exists by which such differentiation might occur. Although it generally is acknowledged that arc (subduction zone) volcanism – the production of andesites – is a principal means by which mantle material is converted to a buoyant state and accreted onto continents, the eventual conversion of that material to a truly granitic composition remains something of a puzzle.

The currently favored picture for granite formation, shown in Figure 16.4, is that partial melting within the continental crust itself produces felsic rocks, such as granites and the granodiorites (intermediate between granites and diorites), which rise to the surface in the form of large intrusive masses called

batholiths, leaving behind a residue in the lower continental crust. However, samples of the lower continental crust are not consistent with being simply a residue of such melting. These *xenoliths* contain trace elements, such as the rare-earths, whose composition is altered by the formation of granite. The abundance pattern seen in these elements in granites versus xenoliths requires that granite formation be more complicated than simple melting and differentiation of the lower continental crust.

One way to explain the pattern is to invoke basaltic magmas beneath the continental crust that rise and plate (or essentially stick to) the bottom of the continental crust. These hot plumes might trigger episodes of deep continental melting and consequent differentiation, while contaminating the lowermost continental crust so as to produce the observed composition of xenoliths. The apparent presence of gabbro in lower continental rock, based on seismic data, is consistent with this idea. Such a model must be tentative at least, because xenoliths may or may not be representative of lower continental rock. More widespread samples of metamorphic rocks called *granulites*, which appear to have formed under high pressure and temperature, may have originated in the lower crust as well; however, there is no consensus on whether these are truly rocks from the lower continental crust.

The difficulty in understanding formation of granitic rock is in large measure a result of the very complex nature of the continents. Unlike the ocean floor, with its simple geology that is erased on 100-million-year timescales, the continents are cumulates of geologic processes stretching over billions of years. To understand how this process began requires examining the nature of rocks from the Archean time of Earth's history.

16.5 Formation of protocontinents in the Archean

There are significant differences between Archean rocks and younger continental materials. In Archean igneous and metamorphic rocks, sodium is more abundant than potassium, in contrast to modern granites, which are potassium rich. Archean volcanic rocks have iron and magnesium content much closer to the mantle values than do Proterozoic and more recent volcanic rocks. They are much less abundant in the rare-earth elements and large-ion elements such as potassium, rubidium, uranium, and thorium than are modern andesitic volcanics. A characteristic of modern continental volcanics and sediments derived therefrom is a depletion of the element europium relative to the other rare-earths; no such depletion is seen in comparable rocks from the Archean. Furthermore, Archean sediments show a large degree of variability in their rare-earth element abundances, unlike the more uniform sediments of later times. The general impression of Archean rocks, then, is of a more chaotic, less regular production process that yields a less extreme fractionation pattern than in the granites of today, and more variability. Further, the continental environment upon which these rocks eroded apparently was less well-developed in terms of extensive stream transport and sorting of sediments: Archean sediments seem poorly processed by water, usually being angular and not well sorted by size compared to modern sediments.

Although several models are offered for the formation of continental masses during the Archean, one particular story stands out that links Archean-type granites to high heat flow from Earth and small continent sizes. Because more accretional heat and undecayed radionuclides were present in the Archean Earth than is the case today, the average heat flow was perhaps double the present-day value. This is an interesting number that corresponds to the rate of heat flowing today from the region of Iceland. Iceland is a minicontinent atop the mid-Atlantic ridge, built up from basaltic magmas and lacking a granitic core. Although its origin in a special event of a hot plume intersecting a mid-ocean ridge is not directly analogous to Archean continental growth, it is a reminder that continents need not (indeed, perhaps cannot) start out as granites.

One possible story, then, for the origin of continents goes this way: the higher heat flow of Earth may have organized the structure of the earliest Archean crust into small plates bounded by hot spots through which basaltic magmas rose. These basalts, building the crust up in selected places, perhaps provided the very first protocontinents poking above the ocean surface. Melting within the basaltic cores of these continents, facilitated by the hot spots and generally hotter nature of the Archean crust, allowed more felsic materials to build up while the residue sank into the lower part of the lithosphere (the rigid, nonconvecting part of the mantle), or flaked off the bottom of the plate into the asthenosphere (flake tectonics).

Eventually, the conditions evolved to the point at which subduction of crust could begin. How and why this happened remains unclear; the contrast in stiffness (viscosity) of lithosphere and asthenosphere may have become appropriate, or perhaps the presence of growing protocontinents on plates forced adjoining plates to sink underneath. The subducting basaltic slab would have been warmer than present-day subducting slabs at comparable depths. For this reason, melting of the subducting Archean slabs could have occurred at much shallower depths, and prior to complete dehydration of the slab. Such premature melting is actually inferred to take place today under the southern Andes mountains of South America, where the subduction zone is so close to the East Pacific Rise (a mid-ocean ridge) that the subducting oceanic crust has not cooled as much as is typical elsewhere around the present Earth (see Figure 9.9 of Chapter 9).

Consideration of the composition of the resulting partial melt suggests that it could be sodium-rich granodiorites and so-called "granitoid" (akin to, but not necessarily mineralogically identical to, granitic) composition consistent with the Archean rock record. These melts would intrude into the basaltic protocontinents, cool, and solidify. As the Archean continents grew in this way, erosion would eventually expose the granitoid and granodioritic rocks and form sediments. The presumably small size of the Archean continents limited the length and size of river systems, accounting for the poor sorting and shaping of sediments. As sodium-rich granites became more massive than the hot-spot basalts, and hot-spot injection of basalts became proportionately less important, melting within the continental mass lessened. Its effect on the sodium-rich granites became minor, in contrast to the modern continental granites whose origin in intracontinent melting leads to strong fractionation of elements.

In this picture, summarized in the bottom panel of Figure 16.4, the Archean crust is derived from two sources: basalts possibly originating in numerous crustal hot spots, and sodium-rich granitoid rocks and granodiorites whose formation is made possible by subduction of warm basaltic crust under marginally buoyant protocontinents. Proponents of this view argue that the rare-earth element patterns in Archean rock are consistent with this model, and it has the advantage of at least one location on the present Earth where an analogous subduction episode is taking place. Other proposals for the origin of sodium-rich granites have been made, for example, through direct melting of mantle material that was previously altered through some earlier melting episode. What is of primary importance here is the notion that continents in effect have bootstrapped their way into the crust. An initial episode of the formation of basaltic protocontinents above hot spots provides the seed for further chemical differentiation within or beneath such continents. Once such differentiation produced felsic, low-density crust, the die was cast for the permanent existence of buoyant continents rising above a denser, surrounding basaltic crust. It is natural to ask whether the evidence for the earliest type of basaltic protocontinents might be found anywhere. It is possible that a significant increase in continental growth near the Archean–Proterozoic boundary erased that evidence forever. Beyond the modern terrestrial analogue of Iceland lies potentially important evidence hidden in the highlands of Venus, to which we return in section 16.9.

16.6 The Archean–Proterozoic transition

As time progressed through the Archean, heat flow from Earth declined, and the area of the surface occupied by the growing continents increased. The lower heat flow may have encouraged an evolution toward larger plates typical of those characterizing Earth's surface today. The change to large-plate tectonics is not recorded directly in the geologic record, but the build up of continental-type crust seems to be, almost defining the transition between the Archean and the Proterozoic. In the early Proterozoic, at different times in different locations, sediments rapidly transition from rare-earth patterns typical of Archean to those of the Proterozoic and more recent times. Potassium-rich granites begin to appear and dominate during this time. The widespread occurrence of uranium deposits in the earliest Proterozoic sediments reflects the increased abundance of large-ion elements typical of potassium-rich granites in the upper crust. The bacterial colonies of stromatolites become much more widespread at 2.5 billion years, reflecting perhaps a larger area of stable continental-shelf territory than available before.

What precipitated this dramatic change in the geochemistry of the continents? Most popular is the view that the growing area of the continents themselves was the cause. As thick, rigid lithospheric crust topped by buoyant continents grew in area, the fraction of mantle plumes under continental crust increased. The insulating effect of thick continental crust and the physical deflection of upwelling plumes by the continental base must have had an increasingly profound effect on the manner in which the crust responded to the heat flow from the interior. Collisions of growing continents on adjoining plates to form larger continental masses exacerbated these effects.

One might speculate that a critical point was reached in the system such that massive melting in the interiors of the continental crust was initiated, triggering large-scale differentiation of Archean continental material and basaltic or mantle material underplating the continents. The increased concentration of radioactive elements in the upper continental crust may have encouraged this process. At what point such massive melting would be triggered, and the details of how the growing continents actually precipitated such a crisis, are not yet understood. In this picture, the potassium-rich granites that are so dominant in the continents of today are secondary granites, a consequence of crustal buoyancy ensured by the Archean generation of sodium-rich granitoid rocks and granodiorites.

Whether the Archean–Proterozoic transition reflects just a change in the composition of the existing continental crust, or a real increase in the volume of continental crust, is controversial. One interpretation of elemental and isotopic data in continental rocks would have the volume of continental material at 10 to 20% of the present value at the onset of the transition some 3.2 billion years ago increasing rapidly to 60% by the close of the Archean at 2.5 billion years (Figure 16.5). However, others have argued that 40% of the present continental volume was already present early in the Archean. The idea of more continental area created earlier in the Archean is attractive in providing a more dramatic perturbing effect on mantle heat flow, so as to initiate large-scale continental melting at the Archean–Proterozoic boundary.

Indeed, some geologists have argued that the Earth has had roughly the same area of continental material since the Hadean, and the balance between continental growth and destruction has been possible through subduction of sediments removed from continents by rivers and transported to the seafloor. Yet a different view puts the start of plate tectonics at only roughly a billion years ago, based on the absence or paucity of certain types of rock indicative of plate collisions prior to roughly a billion years ago. The majority of geologists and geochemists working on continental growth subscribe to the notion that plate tectonics was well underway by the middle part of the Archean. However, the timetable for continental growth remains very uncertain. Would that we could take a peak at an image of the Earth a billion, two billion, or three billion years ago!

Regardless of just how much continent was produced in the Archean, what is not in dispute is the tremendous change in the nature of the granitic rocks: only 7% of the present continental area of Earth is composed of Archean-type rocks, either exposed or buried beneath younger sediments (Figure 16.6). Although some Archean granites might be lurking very deep within continental crust, the consensus view is that much or most of Archean continental material was remelted and differentiated at the end of the Archean.

A final point is that, as the whole of the continental crust was transformed by major melting events, the composition of the lower part of the crust must have become increasingly distinct from the upper. Today, the upper 10 km of the continental crust is a distinct geochemical entity from what lies beneath. The lowermost 20 to 30 km of continental crust is the "other side of the coin" that holds important clues to how potassium-rich, modern granites were produced – clues that can be glimpsed

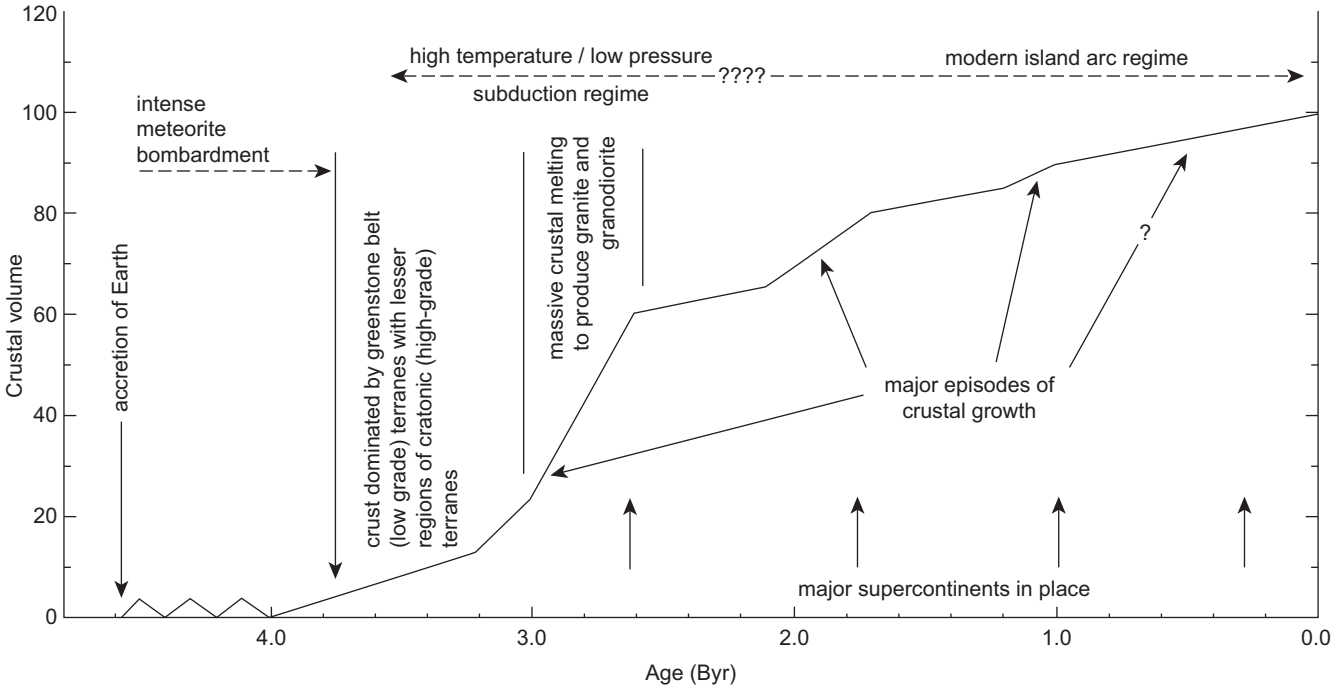


Figure 16.5 One view of the growth of continental volume, relative to its present value, over time. Some other major events in Earth's crustal history are shown, as well as a rough guess as to when modern arc volcanism came into play. Age is in billions of years. From Taylor and McLennan (1995).

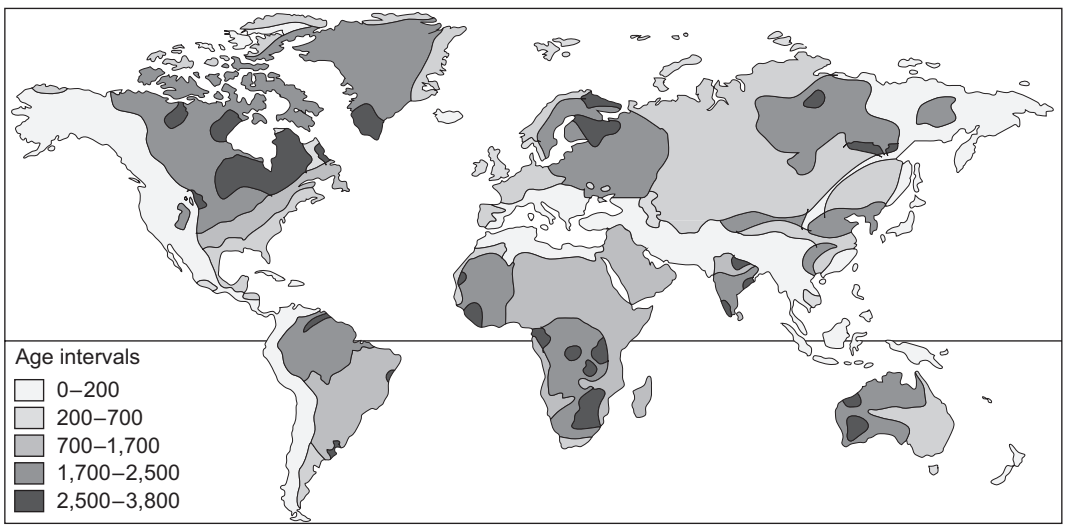


Figure 16.6 Map of Earth showing approximate ages of continental material, in millions of years. Reproduced from Broecker (1985) by permission of Eldigio Press.

only dimly through xenoliths and other bits of the lower crust that are by chance exposed.

16.7 After the Proterozoic: modern plate tectonics

The large-scale build up of continental crust was likely the last step in the development of modern plate tectonics. The continued growth of continents to the present is best explained as beginning at arcs, where buoyant andesite is produced. Motions of converging plates cause continents to ride up over subduction zones, and the andesite is accreted onto the edge of the continent, along with an amalgam of seafloor sediments and small amounts of seafloor crust. As collisions between continents on opposing plates proceed, coastlines disappear, and the *melange* of andesitic and other materials is thrust up in mountain belts along the line of the collision. Metamorphism in these events, along with possible further episodes of crustal melting, continues the geochemical transformation of these rocks.

Geologic evidence suggests that three times in post-Archean Earth history, the continents have joined to form one or two supercontinents. (There is also evidence, albeit very weak, for such a merger at the close of the Archean.) As the continents merged and oceans closed up, the rate of subduction of old ocean crust would have been unusually high, and hence production of andesites at arcs would have increased. The continental collisions themselves could have encouraged further episodes of crustal melting and production of granites, though none as dramatic as that at the close of the Archean. There is weak and controversial evidence that continental crustal growth peaks around times of supercontinent formation.

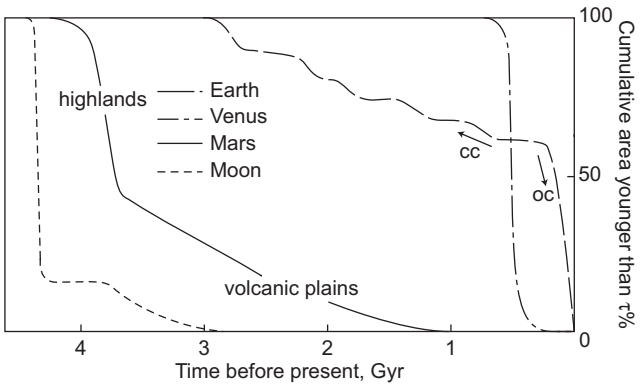


Figure 16.7 Ages of crust on Earth, Venus, Mars, and the Moon, shown as the cumulative amount of area younger than a particular age. The horizontal axis is the age in billions of years before present. Mars is divided roughly into ancient highlands and younger volcanic plains; Earth is even more sharply delineated in terms of age into continental (cc) and oceanic (oc) crust. Adapted from Turcotte (1996).

16.8 Venus: an Earth-sized planet without plate tectonics

Imagine taking hold of Earth in the mid-Archean and stopping once and for all whatever primitive plate tectonics were occurring on the surface. Perhaps a few protocontinents have reached respectable size (close to that of Australia or Antarctica); sodium-rich granite or granodioritic material is beginning to accumulate within the continental cores. Other protocontinents are present, but haven't yet reached the state at which such felsic materials ensure buoyancy; instead, they are rather unsteady basaltic rafts floundering at near-neutral buoyancy in a basaltic crust. In the ocean basins, most of the area of the planet, small-scale subduction zones have been established at the edges of some of the protocontinents, carrying hydrated rocks down to shallow depths where partial melting occurs. As these cease to function (by our imagined interdiction), felsic products of basaltic partial melting stop being produced.

The interrupted planet must still rid itself of heat, and so, basaltic protocontinents continue to form; as they get bigger they eventually founder in the crust. Other hot spots pop up elsewhere and produce new sites of volcanism and growth of plateaus destined eventually to sink. The ocean floor never becomes part of a conveyor belt recycling crust and volatiles; it therefore ages with time along with the rest of the planet. Over billions of years, this ocean floor records the scars of hot-spot formation, foundering of plateaus, abortive attempts at subduction, and asteroid impacts. Over parts of this terrain, large basaltic flows associated with hot spots spread across the surface, renewing portions of it geologically and hiding some of the evidence of past episodes of volcanism.

To find a planet whose surface seems to record such a story, one might turn to Venus. Venus illustrates what happens to an Earth-size and Earth-composition planet on which plate tectonics fails to take hold beyond the early formation of protocontinents. Figure 16.7 compares the distribution of ages of crust on Earth, Venus, Mars, and the Moon. Venus lacks the bimodal ages of continent and ocean floor that Earth possesses. It also lacks the accompanying bimodal height distribution of Earth discussed in Chapter 9, in which the mean elevation of continental crust is well separated from that of oceanic. Instead, a broad range of heights exists on Venus, consistent with: (i) no continuously renewing ocean floor; and (ii) no large-scale production of buoyant granites from the basaltic crust, hence no mechanism to create large continental shields.

Vast areas of volcanic flows are seen in *Magellan* radar images. Magma driven up to the surface by hot spots from the deep interior is generally assumed to be the cause of the volcanic rises seen in the images. Crater statistics suggest that a global resurfacing event, caused by widespread volcanism, occurred between 300 million and 600 million years ago (Chapter 15). An alternative viewpoint is that the Venusian crater record is consistent with a smooth distribution of crustal ages, and that there was not a global-scale catastrophic event in the last half-billion years, but rather, a more continuous resurfacing through regional volcanism. Regardless of the interpretation, Venus lacks the sharp distinction between old continental and young oceanic crust seen on Earth.

The geology of the Venusian surface is as complex as that of Earth, and an adequate description would fill a book. Careful study of the *Magellan* images and other data returned by that spacecraft strongly suggests that Venus does not currently have plate tectonics (Chapter 15) and instead rids itself of internal heat through vertical tectonics. An analogy provided by planetary geologists R. J. Phillips (Washington University) and V. L. Hansen (Southern Methodist University) is that "the crust is not recycled, but instead acts like a rug that locally rips and crumples as a result of relative displacements of domains within the mantle below. The local rips and rumples in the crustal rug of Venus are never far from where the rock that comprises it differentiated at depth. The surface is replenished or repainted with volcanism . . ." (Phillips and Hansen, 1994, p. 648).

Few scientists who study Venus would argue with this view; the disputes are in the details of individual types of surface features. Figure 16.8 provides one view of the origin of major types of Venusian terrain, by Phillips and Hansen. Although others have alternative mechanisms for forming, for example, Ishtar Terra, the general picture that there is little or no horizontal movement is fairly widely accepted.

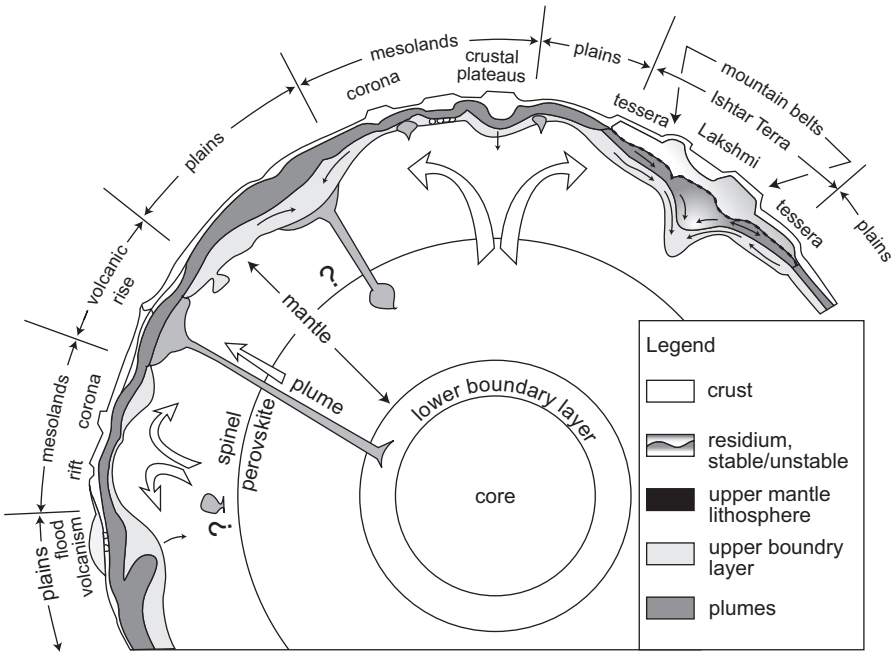


Figure 16.8 One view of the geology of Venus. Mantle plumes rise beneath volcanic regions, and may cause some plains to sag and force crust downward. Other plains may lose crust by delamination or flaking of crust into the mantle. Upwellings are topped by mesoland plateaus, which may be buoyant in crude analogy to Earth's continents. Ishtar is a special region where a particularly buoyant plateau, Lakshmi Planum, is dominating the tectonics. Compare these processes with those for the present and Archean Earths in Figure 16.4. Adapted from Phillips and Hansen (1994) by permission of Annual Review, Inc.

The Venusian mesolands seem the closest analogue to mid-ocean ridges on Earth. They are places where plateaus may be created because of upwelling of warm rock in the mantle. Here, new crust may well be generated. However, the crust does not spread horizontally but piles up vertically. The accumulating material may be slightly buoyant as a result of partial melting but, in the absence of water, the production of granitoid rock of any kind is not expected. If there is no crustal spreading on Venus, then subduction is unlikely to occur there, either. Various features seen in radar imagery have been argued to be the surface expression of subduction on a planet without well-defined plates, but the evidence is not compelling. On a planet without horizontal crustal movement, the loss of crust may come about through *delamination* or flake tectonics, in which material peels off the bottom of the crust and falls into the mantle. Where such delamination is occurring is unclear, but both the plains and the lower parts of Ishtar have been invoked as candidates. On Earth, subduction of cold slabs of crust account for 70% of the heat flow from the mantle. On Venus, delamination may be part of the process by which Venus sheds heat, but it is also possible that much of the heat is transported outward by simple conduction.

Ishtar Terra is a particular enigma. It contains some of the highest terrain on Venus, and displays well-formed mountain belts. It is tempting say that Ishtar is the most developed of the continental masses on the planet, and indeed is the size of Antarctica. However, the origin of Ishtar is controversial, and it has been argued that this unusual part of Venus formed not by plate tectonics but rather as a result of mantle downwelling beneath a buoyant part of the crust. The buoyant zone corresponds to Lakshmi Planum, a broad plateau at the center of the Ishtar mass framed by two mountain ranges. Thus, Ishtar may represent the response of the planet's crust to the presence of a buoyant pseudocontinent, Lakshmi Planum. The geology by which all this happened is likely quite ancient, certainly predating the most recent episode of global volcanism or resurfacing. But it makes Ishtar Terra, and Lakshmi Planum in particular, a high-priority target for any future landed mission to study Venusian geologic processes and understand how such different geologic processes might have operated on an Earth-sized world.

16.9 Water and plate tectonics

An examination of Venus is inconclusive with regard to when it diverged from Earth in terms of geologic styles. Nonetheless, it is tempting to ascribe the difference to the lack of water on Venus. On Earth, modern plate tectonics became possible in the Archean when felsic magmas were produced, allowing protocontinents to be buoyant in the basaltic crust. The precise means by which such magmas form is not important here, because most proposals require that water be carried into the mantle to lower the melting point and alter the nature of the melt products of mantle and basalt. In the absence of such water, buoyant felsic magmas cannot be produced in abundance.

Geochemists I. E. Campbell of Canada and S. R. Taylor of Australia go further. Two decades ago, they pointed out that large amounts of granite must have been produced in the Archean, and this is true regardless of whether by mid-Archean there was 10%

or 40% of the present continental volume. Hence, large amounts of water must have moved through hydrothermal regions in the basaltic ocean crust to hydrate it effectively. It is not sufficient to have had a few lakes here and there on the surface of Earth; a deep ocean is required over much of the Archean planet to ensure large-scale production of granites.

Campbell and Taylor state it eloquently in their 1983 paper: "*Water is essential for the formation of granites and granite, in turn, is essential for the formation of stable continents. The Earth is the only planet with granite and continents because it is the only planet with abundant water*." Venus, having lost its ocean early on, was stopped at the protocontinent stage when, at best, only small amounts of felsic rock could be produced, and the protocontinents could not become permanent buoyant fixtures of the crust.

Even if Venus did not have continent formation, could it have sustained plate tectonics for a long period of geologic time? The extensive resurfacing recorded in *Magellan* radar images makes answering this question difficult. However, subduction may work on Earth because water weakens faults in the lithosphere, allowing this relatively rigid layer of the crust and upper mantle to slide over itself. The lack of water on Venus might have shut down plate tectonics by creating a much stronger lithosphere. Mantle convection would have continued (and probably does today), but underneath a strongly rigid lid that prevents subduction and encourages fixed sites of prodigious volcanism that covers the surface with basaltic lavas.

To fully comprehend the history of tectonics on Venus requires a technologically-challenging program to return to the torrid surface of that planet, to chemically sample the rocks over large areas of the highland regions and plains. The high temperatures of the Venusian surface are very hard on present-day electronics and machinery, but to drill beneath the surface of Lakshmi Planum in search of granites would be to search for that golden nugget that would tell us just how long Venus was a habitable planet (if ever) like the Earth.

One final, speculative point about plate tectonics and Earth's oceans should be made. The mean elevation of the ocean relative to the continents has fluctuated over time, but not by very much. Most of the fluctuation comes from pulses in plate motion: when the continents are merged together, new ocean floor is not produced at many sites, and the ocean crust is unusually cool in the absence of active ridges. A cooler crust is contracted relative to a hot one, and hence the sea level falls relative to the continents. Conversely, when the continents are dispersed and many active ocean ridges exist, the average ocean crust is hotter and expanded, and hence sea level is higher. These sea-level changes have created and destroyed ecological environments on continents, and perhaps have been an important stimulus for the evolution of life.

More controversial and less well understood is whether the sea level has fallen progressively over time from the Archean to the present, as the heat flow from Earth has decreased and plate activity has fallen. The post-Archean geologic record provides no evidence of changes in sea level greater than a few hundred meters in either direction. Either geologic processes have not been such as to force a larger variation, or the volume of water in the oceans adjusts in some way to large changes in the volume of the seafloor crust. This latter possibility is intriguing because the mantle can hold significant amounts of water. Models have been proposed in which the dynamic exchange of water between ocean and mantle, via plate tectonic processes, regulates the volume of ocean water through various feedback mechanisms. As yet these seem speculative.

16.10 Continents, the Moon, and the length of Earth's day

The growth of continental landmasses created obstacles around which the ocean waters flow; at continental edges, ocean waves eroded rocky material and created shallow shelves and beaches. In such regions the paired gravitational pull of the Sun and the Moon produce the high- and low-tide patterns with which we are familiar. As the Moon orbits Earth, the oceans respond to its pull much more than does the solid crust of the Earth. However, the effect of the rising and falling of the oceans is to dissipate some of the energy associated with the tides, in the form of ocean waves and friction along the sea bottom. The net result of this loss of energy is the continuous transfer of angular momentum (Chapter 10) from the rotation of Earth to the orbit of the Moon: the Moon is spiralling outward and Earth's rotation is slowing.

Evidence for this gravitational game of tug-of-war should exist; the length of the day must have been progressively shorter further in the past, and indeed a variety of indicators show this to be the case. The most ancient reliable records stretch back to the Proterozoic in the form of mudstones and sandstones that are stacked in sequences of thicker and thinner layers. These layers are created by the daily changes in the velocity of currents in regions sensitive to tides: the fronts of river deltas, tidal channels, tidal flats, and estuaries. Variations in the daily ebb and flow have a definite relationship to (a) the monthly modulations of high and low tides, as the Moon swings around Earth once every 28 modern days, and (b) the annual cycle of the Earth–Moon system's motion about the Sun. Both the number of days in a lunar orbit and in the year are clearly seen in recent mudstones and sandstones.

Analysis of 900 million year old mudstones and sandstones shows layering implying a different number of days per lunar orbit and per year than at present: the data require the days' length in the late Proterozoic to be only 85% of the length of modern days. Furthermore, by comparing with later mudstones and sandstones, as well as with the modern rate of retreat of the Moon determined by precise measurement of the Earth–Moon distance by laser over 25 years, they conclude that the Moon is moving away from Earth more quickly today than 900 million years ago. The rate of lengthening of the day must also be higher today than during the late Proterozoic. Thus the particular configuration of the modern continents may be serendipitously favorable for the dissipation of tidal energy.

The significance of the existence of continents in the context of the lengthening day is that these landmasses create the environments within which tidal effects are amplified (estuaries, tidal flats, etc.). Absent the continents, tides would still occur, but the dissipation of tidal energy and transfer of Earth's rotational angular momentum to the lunar orbit would be much smaller. Therefore, the outward spiral of the Moon and the lengthening

of the day was likely a much more gradual affair before the Archean–Proterozoic growth of continents. What the day length was 3 billion years ago is not known, but much of the lengthening may have occurred in the last half of Earth's history. For those of us who find the length of the day much too short, there is at least some comfort in the notion that, absent tides, it would be much shorter.

16.11 Entree to the modern world

From a planetary perspective, the shift to a fully modern plate tectonic mode of crustal heat loss by 2.5 billion years ago represents a key departure of Earth's history from that of Mars and Venus. No more significant geologic change has happened to Earth up to the present. From the standpoint of life, the growth of continents opened up whole new places to live, but it would require another 2 billion years for life to take full advantage of the vast spaces of exposed land.

As the Proterozoic eon began, increasing amounts of photosynthesis, reflecting the growing abundance of life, began to alter the composition of the atmosphere toward an oxygen-rich state. This in turn allowed a profound alteration in the nature of cellular life that was the prerequisite for the kinds of continental ecosystems that we see today. How the oxygen revolution came about, and its implications for life, are the subject of Chapter 17.

Summary

Earth is geologically distinct from its neighboring planets Mars and Venus in having a significant amount of crustal rock with a so-called granitic composition, that is, rich in sodium and potassium, and poor in iron and magnesium, compared to basalts. But even basaltic rock is very different from the building blocks out of which the Earth formed, represented approximately by the composition of chondritic meteorites. Iron is poorly represented in mantle rocks compared to chondrites, pointing to the separation of iron from the mantle to form a core early in the history of the Earth. But the formation of basalt from mantle rock is a further geochemical evolution, a consequence of the effect of pressure on melting of rock, such that basalt is a "partial-melt" product of the mantle. Granitic rock is even more extreme in composition, and is the result of further cycles of chemical refinement that remain poorly understood: the effect of water on partial melting, and melting of rock in the cores of continents, play a role. The formation of continents was a bootstrapping process over time, beginning with small cores of basaltic material that evolved chemically under the action of subduction, growing and becoming progressively more granitic in composition. How this happened is not well understood, because the geochemical record of when subduction – indeed when modern plate tectonics – began is not easy to interpret. The high heat flow in the Archean could have allowed subducting slabs to melt, rather than dehydrate as they do at present – hence leading directly to the formation of rocks with a granite-like composition. The pace of growth of the continents is debated, but there is compelling evidence that the Archean–Proterozoic boundary in the geologic record might have been marked by a significant increase in the growth of continents and the establishment of the modern style of plate tectonics. Venus provides a potential place where Archean-style plate tectonics might be studied, since if plate tectonics began at all there, it likely ended early in Venus' history. With the loss of water, subduction slowed or stopped, the production of granites became impossible, and the planet shifted to a different style of geology dominated by basaltic volcanism. Or so goes the story: to test it will require sampling surface rocks or even drilling beneath the basaltic veneer of our sister planet.

Questions

1. Suppose Earth had remained a waterworld with few continents. How would this have affected the evolution of life, recycling of carbon dioxide, and Earth–Moon orbital evolution?
2. What definitive chemical tests are required on Venus to determine that plate tectonics has not operated there for billions of years?
3. An alternative to the plate tectonics model offered in the 1960s was the so-called oceanization of continental crust: a kind of vertical tectonics in which ocean basics were created by mixing of crust and mantle. Given our knowledge of the chemistry of basalts and granites, argue against this model.

4. Suppose the melting point of mantle rock were to decrease with increasing pressure. On the diagram of Figure 16.2 draw this case and explain under what conditions melting occurs.
5. Make a list of the aspects of plate tectonics, including the differentiation of the various rock types, that depend on the presence of liquid water. Considering each of these effects one at a time, how would the Earth's geologic evolution change in the absence of liquid water?
6. Speculate on the nature of plate tectonics and crustal evolution on a rocky planet more massive than the Earth, with higher heat flow, possibly thinner crust, etc.

General reading

Condie, K. C. 2005. *Earth as an Evolving Planetary System*. Elsevier, Amsterdam.

Gargaud, M., Claeys, P., Lopez-Garcia, P. *et al.* eds. 2006. *From Suns to Life: A Chronological Approach to the History of Life on Earth*. Springer, Dordrecht.

References

Broecker, W. 1985. *How to Build a Habitable Planet*. Eldigio Press, New York.

Campbell, I. H., and Taylor, S. R. 1983. No water, no granites – no oceans, no continents. *Geophysical Research Letters* **10**, 1061–64.

Drake, M. J. and Righter, K. 2002. Determining the composition of the Earth. *Nature* **416**, 39–44.

Harrison, T. M. 2009. The Hadean crust: evidence from > 4 Ga zircons. *Annual Review of Earth and Planetary Science* **37**, 479–505.

Kasting, J. F. and Holm, N. G. 1992. What determines the volume of the oceans? *Earth and Planetary Science Letters* **109**, 507–15.

Kröner, A. 1985. Evolution of the Archean continental crust. *Annual Review of Earth and Planetary Sciences* **13**, 49–74.

Kröner, A. and Layer, P. W. 1992. Crust formation and plate motion in the early Archean. *Science* **256**, 1405–11.

Mason, S. F. 1991. *Chemical Evolution*. Clarendon Press, Oxford.

Phillips, R. J. and Hansen, V. L. 1994. Tectonic and magmatic evolution of Venus. *Annual Review of Earth and Planetary Sciences* **22**, 597–654.

Press, F. and Siever, R. 1978. *Earth*. W. H. Freeman and Company, San Francisco.

Rogers, J. J. W. 1993. *A History of the Earth*. Cambridge University Press, Cambridge, UK.

Roillinson, H. 2007. When did plate tectonics begin? *Geology Today* **23**, 186–91.

Sonnett, C. P., Kvale, E. P., Zakharan, A., Chan, M. A., and Demko, T. M. 1996. Late Proterozoic and Paleozoic tides, retreat of the moon and rotation of the Earth. *Science* **273**, 100–104. Corrigenda *Science* **273**, 1325 and *Science* **274**, 1065.

Taylor, S. R., and McLennan, S. M. 1995. The geochemical evolution of the continental crust. *Reviews of Geophysics* **33**, 241–65.

Turcotte, D. L. 1995. How does Venus lose heat? *Journal of Geophysical Research* **100**, 16931–40.

Turcotte, D. L. 1996. Magellan and comparative planetology. *Journal of Geophysical Research* **101**, 4765–73.

17

The oxygen revolution

Introduction

Perhaps the most fundamental shift in the evolution of Earth's surface and atmosphere was the oxygen "revolution," an event stretching over the Proterozoic eon when molecular oxygen levels in the atmosphere rose and carbon dioxide levels decreased. (Hereinafter, for brevity, we refer to molecular oxygen, which is O_2, simply as oxygen.) In consequence, the fundamental chemical nature of the atmosphere and its interactions with life changed drastically. Life was responsible for, or at least helped to, precipitate the drastic increase in oxygen levels and, as a result, was set on a radical new course. Earth's atmosphere today is not the sedate, relatively unreactive carbon dioxide atmosphere as on Mars and Venus. Instead, it is an atmosphere far from equilibrium, held in a precarious chemical state by the biosphere. As Margulis and Sagan (1986) express it, the modern biosphere hums "with the thrill and danger of free oxygen."

In this chapter we explore how this change came about on the Proterozoic Earth, by first examining the present-day oxygen cycle and the evidence in the rock record for an oxygen-poor Archean and early Proterozoic Earth. We then consider a model that, although approximate and based on mechanisms that are still debated, illustrates very well how the change might have taken place. Such models often have critical utility in science, in that they point the way toward new observations and investigations that will yield deeper insight into a particular process (even while proving the model itself to be incomplete or incorrect).

17.1 The modern oxygen cycle

Figure 17.1 shows the sources and losses (sinks) of oxygen on Earth today. The total oxygen in the atmosphere today is roughly 6×10^{17} kilograms and is held in balance by production (gain) and loss processes, the importance of which may have varied on geologic timescales. (Some readers may find it helpful at this point to review the discussion of scientific notation in Chapter 1.) Here we outline the most important gain and loss processes. We give rates only to the nearest order of magnitude; this is good enough for our purposes, and in many cases the uncertainties do not justify any higher accuracy.

1. *Photochemistry and escape of hydrogen to space*. The absorption of ultraviolet photons from the Sun by water (H_2O) causes the molecule to break up, forming hydrogen and oxygen. The hydrogen can escape from the atmosphere, preventing recombination. The oxygen left behind makes molecular oxygen (O_2) and ozone (O_3). The rate of oxygen production is 10^8 kilograms per year (abbreviated as kg/yr).
2. *Weathering of rock*. Oxygen and carbon dioxide in the atmosphere, with the help of water, attack minerals in the rock to make new compounds, which precipitate out as sediments (Figure 17.2). In the case of oxygen, which attacks the iron in the rock, the process is akin to rusting. Estimating the rate of this process is not easy because it depends on how rapidly the weathered products are transported to the ocean by river systems, but is approximately -10^{11} kg/yr. The negative sign indicates that this is a loss process.
3. *Volcanism*. Volcanoes on land and the ocean floor emit reduced gases, such as carbon monoxide and sulfur compounds, that strongly tend to combine with oxygen in the atmosphere. The resulting rate of oxygen loss is about one-third the rate caused by weathering. Volcanoes also emit water vapor, which, through photochemistry and loss of hydrogen, produces oxygen as described above.
4. *Photosynthesis*. Carbon dioxide is removed from the atmosphere by plants and bacteria and molecular oxygen is produced. The rate of oxygen production from photosynthesis is 10^{14} kg/yr.

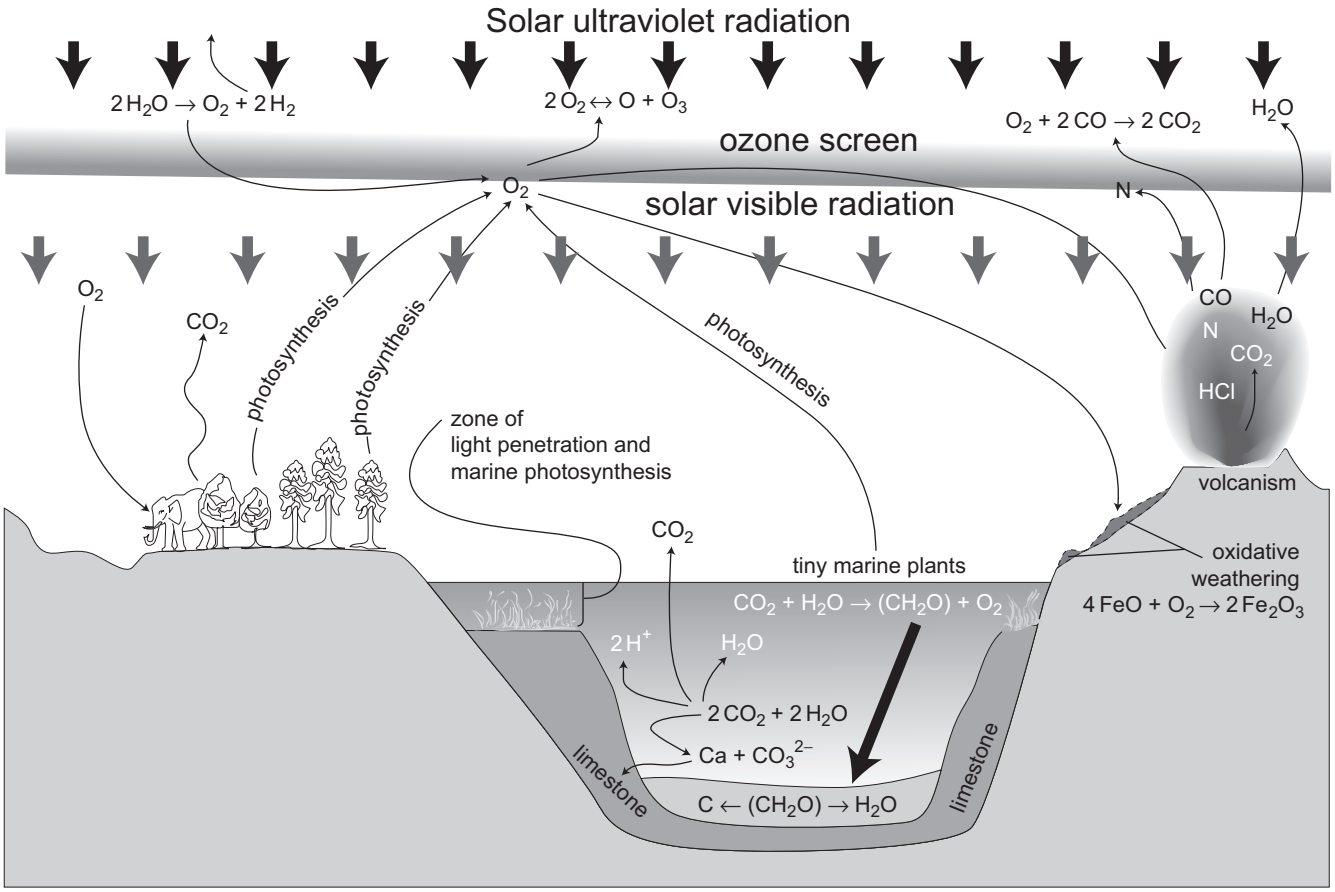


Figure 17.1 Oxygen cycle on Earth today, showing processes that are significant in producing or destroying oxygen. Chemical reactions involving oxygen are summarized; the actual chemistry involves many more steps than the equations on the figure show. Based on Cloud (1988).

5. *Respiration and decay*. These two processes are, with respect to oxygen, the reverse of photosynthesis. Oxygen is taken up from the air by animals, plants, and certain bacteria and combined with sugars or other organic compounds to generate energy (along with carbon dioxide, water, and other products). Decay refers to organic matter that is no longer living but is consumed by microorganisms, with a net loss of oxygen, to generate energy as described in Chapter 12. The observation that the present level of molecular oxygen is approximately constant and the fact that respiration would deplete the atmosphere of oxygen in 6,000 years imply that respiration/decay is in balance with photosynthesis. If the number of plants were to suddenly increase, enhancing the level of oxygen, surface decay processes would speed up as the bacterial population grew to take advantage of the additional oxygen. Note that only three-fourths of the surface organic reservoir in contact with the atmosphere today is living. The total rate of oxygen loss from these processes is -10^{14} kg/yr.
6. *Burial of carbon from organisms*. Computations show that, on average, the remains of dead organisms lie on the surface, in contact with the atmosphere, for several decades or more. We refer to this carbon, which is relatively rich in hydrogen and tends to soak up oxygen, as *reduced carbon*. (Some workers in the field refer to this material as organic carbon, but we have previously used the term organic in other ways.) The primary means of burial of the reduced carbon is deposition in continental and oceanic sediments, which breaks the contact with the atmosphere and allows the carbon to be preserved. Because the buried carbon is no longer available to soak up oxygen, the net result is that oxygen is added to the atmosphere over time. The effective rate of oxygen production is 10^{11} kg/yr.
7. *Recycling of buried sediments*. As discussed in Chapter 14, ocean-floor sediments containing trapped carbon are recycled through the upper mantle by plate tectonics. The cycling time is roughly 100 million to 200 million years. The result is the re-emergence of reduced carbon at the surface, a net source of carbon dioxide and sink of atmospheric oxygen. The amount of oxygen loss is somewhat less than the production rate associated with the sedimentary burial of reduced carbon given above.
8. *Fossil fuel combustion*. This is an artificial form of weathering, caused by human burning of oil, coal, natural gas, and other fossil fuels extracted from deep sedimentary layers. The rate of oxygen loss, -10^{12} kg/yr, is much larger than for natural weathering. It will be short lived on geologic timescales because such burning began in earnest during the

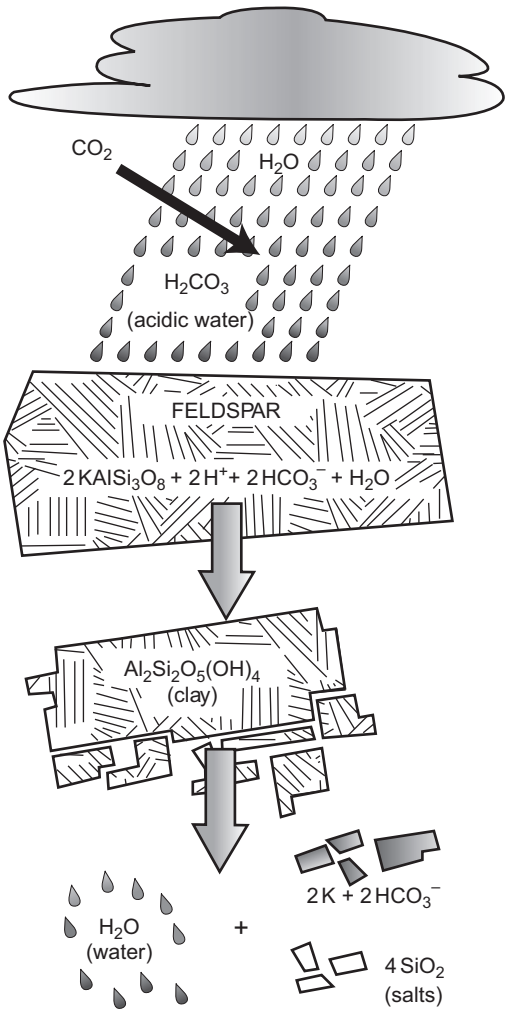


Figure 17.2 Example of the weathering of rock: in this particular case through the action of water and carbon dioxide.

seventeenth century Industrial Revolution and will cease as we deplete these resources within the next century or so (Chapter 23).

17.2 The balance of oxygen with and without life

A look at the numbers given above shows that photosynthesis is the most important source of oxygen. Respiration/decay must be the primary balancing mechanism for losing oxygen because none of the geologic processes are speedy enough to balance photosynthesis. What was the situation before life became abundant? We can compare the most important nonbiological processes for gaining and losing oxygen, which are photochemistry at 10^8 kg/yr and weathering at -10^{11} kg/yr.

Clearly, photochemistry cannot generate oxygen quickly enough to keep pace with the destruction by weathering and volcanism. Because there are currently 6×10^{17} kg of oxygen in the atmosphere, weathering and volcanism could destroy almost all of the oxygen in the atmosphere in 6 million years at its present rate (6×10^{17} kg/10^{11} kg/yr $= 6 \times 10^6$ years). So, if we consider the roughly one-billion-year period before the emergence of photosynthesizing life-forms, it becomes a sensible notion that free molecular oxygen must have been very scarce in that atmosphere in the presence of weathering and volcanism and in the absence of photosynthesizing life.

17.3 Limits on oxygen levels on early Earth

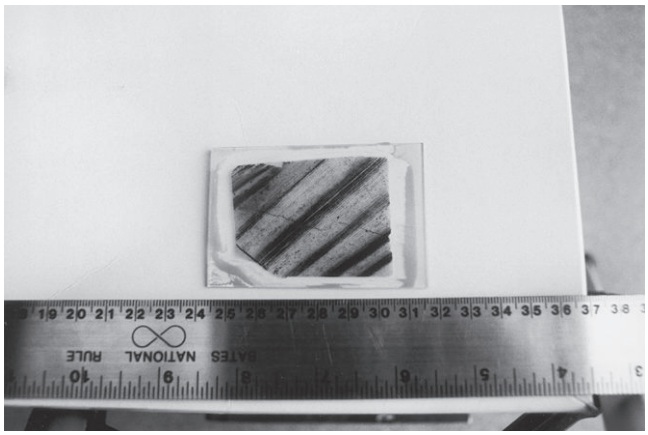
Photosynthesis in the time of the first stromatolites that have been found in the fossil record (3.5 billion to 3 billion years ago) was probably not widespread; consequently, the rate of oxygen production was less than it is today. The early oxygen was likely "soaked up" by weathering and by vigorous volcanic activity. Evidence for the early oxygen abundance being low, increasing significantly only after 3 billion years ago, is to be found in a number of parts of the rock record, the most important of which are outlined below.

17.3.1 Minerals unstable in the presence of oxygen

The early continental rock record, up to about 2.7 billion years ago, shows fragments of rock containing the minerals pyrite and uraninite. Pyrite is FeS_2 and, in the presence of oxygen, would react such that the iron combines with some of the oxygen to form iron oxides. Uraninite is UO_2 and uranium tends also to form other oxides. Note that the presence of significant amounts of uranium in the crust was, as discussed in Chapter 16, a consequence of the partial melting process that led to continent formation. Here, the particular chemical form in which uranium exists in ancient rock deposits tells us something about the amount of free oxygen that could have existed in the atmosphere at the time that rock was first exposed at the surface. Had there been significant amounts of oxygen in the atmosphere 2.7 billion years ago, the uraninite and pyrite fragments would have been chemically altered through exposure to the air. (Subsequent burial of the rock, until more recent extraction, ensured that the more modern oxygen-rich atmosphere had no effect; undoubtedly other uraninite deposits have been destroyed over time.) Pyrite and particularly uraninite suggest that the Archean and early Proterozoic atmospheres had very little molecular oxygen.

17.3.2 Banded iron formation

The banded iron formations (BIFs) occur commonly among sedimentary rocks dated in the 2-billion- to 3-billion-year-old range, with a few older examples. They are extremely rare or nonexistent in younger rocks, the exception being some dated at 750 million years ago, possibly corresponding to a deep period of near-global glaciation. They consist of alternating dark bands containing up to 30% iron, and light bands made of silica (chert) (Figure 17.3). These bands retain their distinctiveness over vast horizontal lengths of hundreds of kilometers. To form such bands



(a)

(b)

Figure 17.3 Banded iron formation rocks from (a) the Proterozoic and (b) the late Archean. Panel (a) is a thin section of the Proterozoic rock mounted on a glass plate. Scale is in centimeters.

required that iron be dissolved in ocean water, then deposited repeatedly on top of layers of accumulating chert on the seabed. The sediment then was compressed, forming over time a hard rock. Banded iron formations are found essentially on all continents, and make up more than 90% of the world's commercial iron supply.

The curiosity about BIFs lies in the need to dissolve iron in water during their formation – it cannot happen under today's oxygen-rich atmospheric composition. The form of iron that dissolves in water (FeO) is called ferrous iron. Oxygen in the atmosphere today is partly dissolved in the ocean, and can then combine with the ferrous iron to make ferric iron, Fe_2O_3. Ferric iron is more oxidized than ferrous – that is, the element iron has bonded with more oxygen atoms than in the ferrous state: three oxygen atoms for every two iron atoms, instead of one to one. The ferric iron immediately precipitates out of the water and falls to the seafloor as iron-rich particles.

To maintain iron in the ferrous form, and hence soluble in the ocean, required an atmosphere that was relatively oxygen free. This sets limits on the amount of oxygen in the late Archean and early Proterozoic, 2 billion to 3 billion years ago, at a few percent of the present-day value or less. However, a mystery still remains: given a mechanism for dissolving the iron in the water, the production of BIFs then requires periodic precipitation of the iron out of the water.

The problem is unsolved, but one idea goes as follows: dissolved iron was contained in deep-ocean water near active vent sites. These iron-rich waters would spread by mixing over large areas of the ocean. Upwelling of this water to the near-surface brought it into contact with regions in which cyanobacteria existed, and hence photosynthesis took place. At certain seasons of the year, or perhaps stimulated by sufficiently large amounts of dissolved iron, the bacteria would increase their photosynthetic output of oxygen. Beyond a certain point, the oxygen produced by the bacteria would combine with ferrous iron to make ferric iron, which would precipitate out. In shallow ocean areas, the iron would precipitate out onto chert layers, forming one set of alternating bands. As the photosynthesis slowed again, oxygen levels in the water would decrease, iron could be stable again in the dissolved ferrous form, and the cycle would repeat. These layered sediments, over time, eventually would be compacted and lithified.

The cyanobacteria were not the only life-forms participating in this process. Certain other kinds of "rusting" bacteria take oxygen from the surrounding environment and combine it with iron, creating stored energy usable for their life processes. This would have assisted the process of iron precipitation. In the summer, when cyanobacteria were active in producing oxygen, the rusting bacteria would have been more abundant and extracted more iron from seawater. In the winter, with less oxygen produced by cyanobacteria, biological deposition of iron-bearing sediments would have slowed or stopped. The variation from place to place in the width of the iron bands – from micrometers to meters – suggests that oxygen levels fluctuated on seasonal and longer (perhaps decades or more) timescales in different places at different times.

An explosion in the production rate of BIFs in the 2.2 billion- to 1.8-billion-year time frame suggests that oxygen levels worldwide had by then reached a threshold at which variations in photosynthetic activity modulated the precipitation of iron from oceans. However, some rare cases of BIFs occur prior to 3 billion years ago (perhaps as early as 3.86 billion years), when worldwide oxygen levels were very low. These oldest BIFs hint that, in localized areas, some form of photosynthesis intensive enough to produce significant quantities of oxygen might have occurred. Alternatively, mechanisms have been suggested by which molecular oxygen produced by atmospheric photochemistry might periodically have been concentrated in localized environments, but they remain speculative.

17.3.3 Redbeds

Beginning about 2 billion years ago and extending to recent times, sediments appear in the rock record that require oxygen for their formation. These *redbeds* form when iron is weathered out of rock in the presence of oxygen. The threshold amounts of oxygen that are required to make redbeds are significant but still small enough to permit BIFs to exist; the two overlap in the geologic record by several hundred million years.

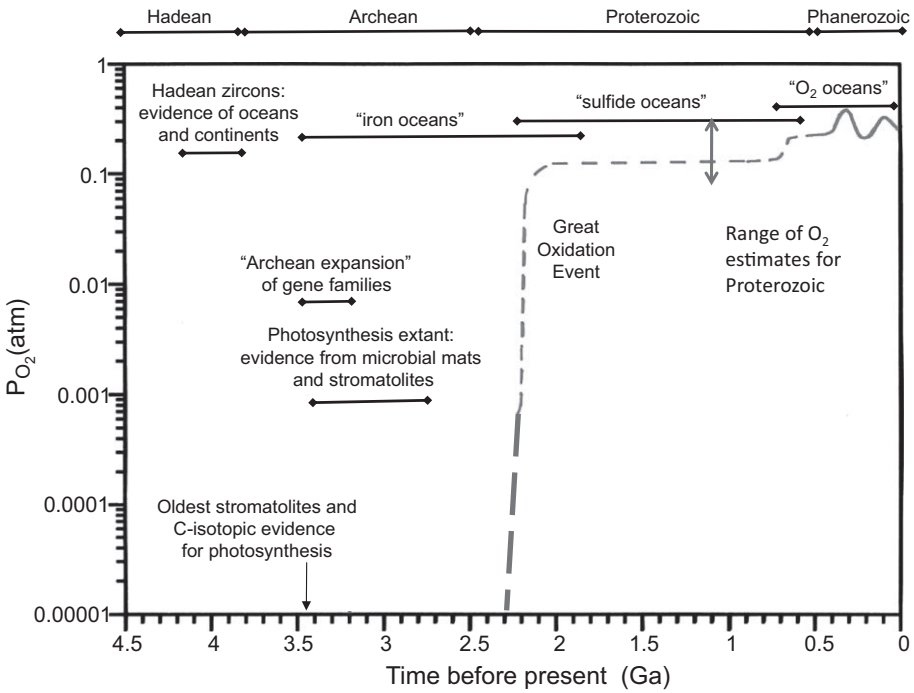


Figure 17.4 History of oxygen abundance in the atmosphere of Earth, assembled from diverse pieces of evidence described in the text. In most cases, the constraints are weak, or provide only upper and lower limits. Various geologic events, are also listed. The "iron" ocean refers to the ocean when oxygen levels were low but fluctuating such that BIFs could form. The sulfide ocean refers to a model proposed by Canfield, in which the rise of oxygen in the oceans led to a period in which the oceans were sulfide rich.

17.3.4 Fossils of aerobic organisms

Before the advent of free oxygen, organisms produced energy for biochemical processes in a number of ways. The most familiar process, one still in operation in oxygen-poor (anaerobic) environments, is fermentation (Chapter 12). Here, sugar is converted to ethanol and other molecules, with release of energy. The energy is stored in a biological molecule containing phosphate bonds, called adenosine triphosphate (ATP). One molecule of sugar makes enough energy to be stored as two molecules of ATP.

Respiration, as discussed in Chapter 12, uses oxygen to convert sugars to carbon dioxide, water, other products, and a great deal of energy. Respiration can produce up to 36 ATP molecules from one sugar molecule. This tremendous boost in bioenergetic efficiency allowed explosive growth in the number of forms of cyanobacteria in the Proterozoic eon, and later enabled complex cellular life (eukaryotes) and multicellular eukaryotic life (plants and animals).

The times at which these biological events appear in the fossil record in rocks and the known biochemical requirements for oxygen among such species today allow the increase in oxygen in the atmosphere to be tracked. The late and relatively rapid appearance of large, complex, multicellular animals only 550 million years ago suggests that oxygen levels may have remained well below the present value (perhaps 10 times less) until then. The possibility of a dip in the molecular oxygen abundance, associated with a sharp decline in plant life associated with the so-called "neoproterozioc glaciation" 750 million years ago, is suggested by the appearance of BIFs dating to that time. The evidence for charcoal in the fossil record of the past 100 million to 200 million years implies forests capable of undergoing combustion (burning); this requires oxygen levels close to those at present (13% compared to the present value of 21%).

17.4 History of the rise of oxygen

With the evidence for an early time of little or no atmospheric oxygen and a significant increase beginning in the Proterozoic eon, we can put together a chart (Figure 17.4) of the amount of oxygen in the atmosphere over Earth's history. The chart is rough, showing much uncertainty in the actual levels, but the general nature of the conclusion is clear: before the start of the Proterozoic, oxygen was a very minor component of the Earth's atmosphere.

How did the change come about? The clues are present in the evidence described here, but must be assembled carefully into a working hypothesis. Such a hypothesis ought to explain the physical evidence in terms of the processes that occurred over time to generate the oxygen-rich atmosphere. We next consider one possible model for the growth of oxygen.

17.5 Balance between oxygen loss and gain

Earlier in the chapter we considered present-day rates of oxygen production and loss. These rates were different during the

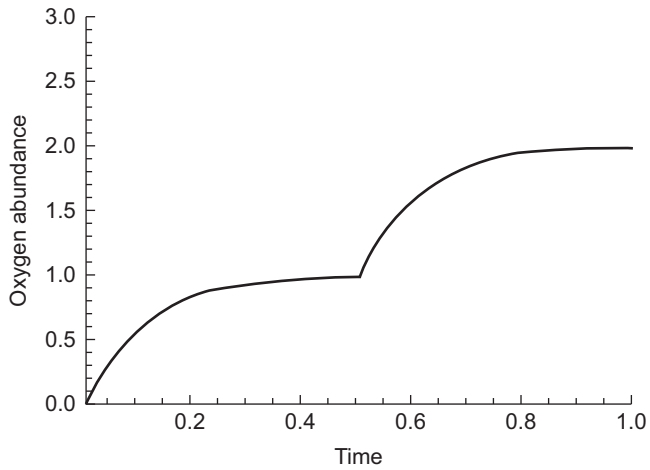


Figure 17.5 Example curve of oxygen abundance in the face of changes in production or loss processes. The horizontal axis is time, and the vertical axis is oxygen abundance (both in arbitrary units). As the oxygen abundance approaches a constant value, reflecting a balance between production and loss, a suddenly increased production rate causes a jump in abundance followed by leveling off at a new, higher value. The increased production rate might be due to a novel source, or simply an increase in production from an established source of oxygen.

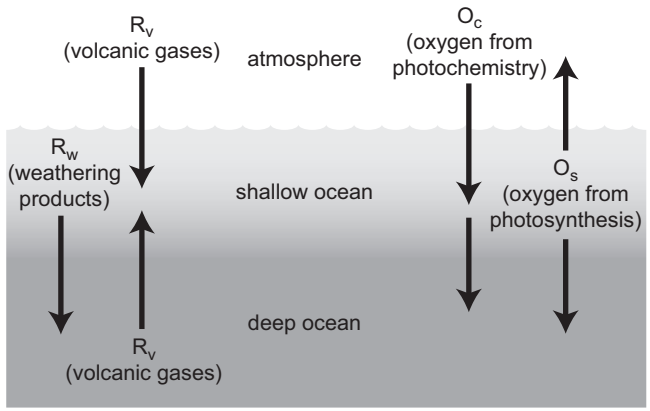


Figure 17.6 Box model of Earth used to understand the growth of oxygen. Three components of Earth are atmosphere, shallow ocean, and deep ocean. Sources of oxygen are labeled "O," and sinks are labeled "R," with subscripts to distinguish among them. Weathering products include not only sediments but also reduced carbon from dead organisms, which, left exposed to the atmosphere, can soak up oxygen. Based on the model of Kasting (1991).

Archean epoch compared to the present. Biological processes were not nearly as important then, and respiration in the absence of significant amounts of oxygen must have been negligible or nonexistent. Recycling of crust and mantle may have been much more rapid in the Archean than at present, leading to a greater rate of volcanism at that time. This led to a higher flux of reduced gases into the atmosphere, which would have soak up oxygen at a higher rate.

The change in amount of oxygen in the atmosphere per unit time is simply the rate of production minus the rate of loss. In figuring out how production and loss work to produce a particular amount of oxygen, an important fact is the following: the loss processes depend on how much oxygen is available, but the production processes usually do not. If we start with zero oxygen, there can be no loss of oxygen from weathering or volcanic gases. As more oxygen is produced by photochemistry, more can be lost by weathering and volcanic gases. A graph of the amount of oxygen as a function of time will then look something like Figure 17.5.

As any new source of oxygen arises, loss rates (proportional to the oxygen abundance) increase, until a new steady state is reached, characterized by a constant or only slowly varying oxygen abundance. Alternatively, some loss processes might saturate as the oxygen abundance rises; this would effectively increase the rate at which the oxygen abundance grows. The sinks of oxygen on the early Earth must eventually have been overwhelmed by increasing rates of oxygen production.

17.6 Reservoirs of oxygen and reduced gases

The situation on the early Earth is best summarized by considering reservoirs of oxygen and the substances that can soak up oxygen, via weathering, volcanism, or organic matter from life-forms that have died but have not been deeply buried in sediments. We will simply call these substances *reducing compounds*, meaning elements or molecules that like to combine with oxygen. A simplified model of Earth as just atmosphere and ocean is sketched in Figure 17.6.

We ignore the continents because, even though volcanoes may be on land or sea, and weathering processes start out on land, the "action" ends up being in the ocean or atmosphere. Most of the continental weathering products end up in the ocean, and the volcanic gases are present in the atmosphere or dissolved in the ocean. We must distinguish between the deep ocean, which has slow, limited contact with the atmosphere, and the shallow upper part of the ocean, where photosynthesis takes place (because some sunlight is present) and gases are exchanged with the atmosphere. Included in the shallow part of the ocean are rivers and lakes.

Furthermore, we do not consider the variation from place to place in oxygen content, only the difference between the three environments – atmosphere (top), shallow ocean (middle), and deep ocean (bottom). This is called a *one-dimensional model*; it is useful in understanding many physical situations because of its simplicity. Obviously, such models cannot explain fine details, and may miss important processes that occur or vary from one place to another, but our information on oxygen abundance on the early Earth is so limited that this simple model has great utility. A reminder of its limitations is the presence of BIFs in the Archean, which indicates oxygen variations from one location to another on Earth.

Over time, we distinguish between three states of each of the reservoirs:

1. *Reducing*. This means that the reservoir has so little oxygen that minerals such as uraninite will be stable, and iron can remain in solution in the ocean water, which is required to produce BIFs.

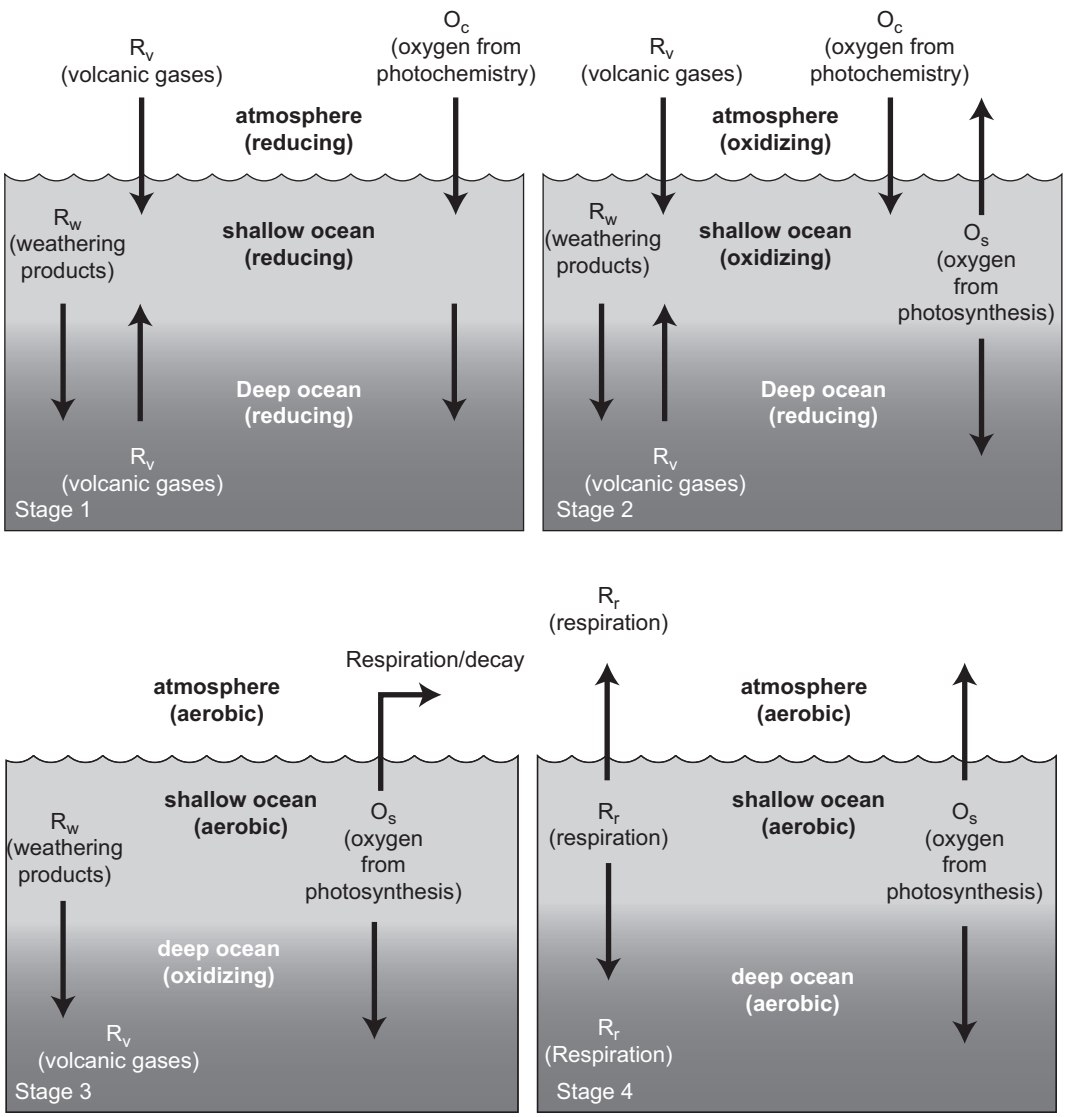


Figure 17.7 Four stages in the history of oxygen on Earth, distinguished by the oxidation state of the three major oxygen reservoirs considered in the model. Important production and loss processes in each stage are shown. Based on the model of Kasting (1991).

2. *Oxidizing*. Here, the reservoir has enough oxygen to make minerals such as uraninite unstable, and to prevent iron from staying dissolved in seawater. However, not enough oxygen is available to sustain aerobic respiration.
3. *Aerobic*. Enough oxygen is present to allow aerobic respiration to occur.

With this model we can map the history of oxygen in the four stages illustrated in Figure 17.7. All oxygen abundances are listed in fractions of the present atmospheric level (PAL).

17.7 History of oxygen on Earth

17.7.1 Stage 1

Once water is established on Earth (Hadean–Archean boundary), photochemistry begins to produce oxygen. The oxygen levels off as production is balanced by weathering and volcanism. Oxygen in the atmosphere ranges between 10^{-5} and 10^{-13} PAL. The range is based on detailed calculations by Pavlov and Kasting that consider the record in Archean sediments of the trend in abundance of the four stable isotopes of sulfur. The trend does not depend on the differences in mass between the isotopes – so-called "mass independent fractionation"). It suggests that a variety of different sulfur compounds, with different oxidation states (that is, different amounts of hydrogen and oxygen), were produced photochemically in the atmosphere and then preserved during their removal into sediments. Even trace amounts of oxygen exceeding parts per million, they argue, would have forced a more uniform set of oxidation states and led to a very different pattern of isotopic ratios. This reducing environment would also preserve uraninite. Banded iron formations from this time occur, and were formed either as solar ultraviolet radiation (which reached Earth's surface in the

absence of an ozone shield) oxidized iron in water, or in localized environments where photosynthesizing organisms such as cyanobacteria were concentrated.

17.7.2 Stage 2

The spread of oxygen-producing photosynthesizing organisms around the planet initiated a new source of oxygen. At first, aerobic photosynthesizers would have been restricted in geographic extent, perhaps because they were not very tolerant to the oxygen they produced (limiting them to environments with strong oxygen sinks), or perhaps due to competition with anaerobic photosynthesizers that may have preceded them. The geologic data suggest that oxygen in the atmosphere jumped to between 10^{-2} PAL and 10^{-1} PAL around 2.2 billion to 2 billion years ago, enough to be considered oxidizing for most minerals. The geologic record for this time shows an overlap between the occurrence of BIFs and redbeds. The oxygen abundance in stage 2 is small enough that the deep ocean could have remained oxygen poor (reducing), whereas the upper ocean, where photosynthesis took place, would have been oxidizing. Under such circumstances, deep-ocean water containing dissolved iron may have slowly circulated up to the surface, where it encountered oxygen-rich conditions and precipitated out iron, forming BIFs.

The steep rise in oxygen during this time period has prompted speculations on mechanisms beyond increased photosynthesis to pump up the atmospheric oxygen level. Geologic evidence suggests that around this time a number of small continents collided to form the first *supercontinent*, a process to be repeated again and again in more recent history (Chapter 19). NASA Ames scientist David Des Marais suggests that the assemblage of continental fragments into larger masses had as a side effect the increasing rate of burial of dead organisms (what we have called reduced carbon). The heightened burial rate occurred both directly in the extensive continental interiors and on the seafloor; rates on the seafloor were enhanced as large mountain ranges, built up on the colliding continents, sped the delivery of sediments to the sea. With much smaller amounts of reduced carbon exposed to the atmosphere, less absorption of oxygen by these compounds could occur; in effect an important sink of oxygen was eliminated. Alternatively, as argued by Pennsylvania State University scientist Jim Kasting, by this point in Earth's history large amounts of ocean water were mixed into the mantle by plate tectonics (equivalent perhaps to half the volume of the present oceans). This process would gradually have turned the mantle from a reducing to an oxidizing chemical state, such that volcanic gases emanating from the mantle became progressively less effective in soaking up atmospheric oxygen. The decreased importance of volcanism as a sink combined with increasing rates of oxygen production from photosynthesis led, in this picture, to a steep increase in abundance of atmospheric oxygen.

17.7.3 Stage 3

As photosynthesizing organisms proliferated, the oxygen content of the atmosphere increased. Several factors may have limited the rate of this increase. Environmental factors, such as near-global glaciation episodes, could have reduced the available surface area of liquid water on the Earth, dramatically reducing the population of oxygen-producing photosynthesizers for hundreds of millions of years. Also, rising oxygen levels might have provide a challenge to the defensive mechanisms against oxygen-generating free radicals within the organisms themselves. It is not too much of a fantasy to imagine that, had the evolving genome not been sufficiently flexible or inventive, all such photosynthesizers might have poisoned themselves to or beyond the brink of extinction, forever limiting the amount of oxygen in the atmosphere. Happily this was not the case, and by 1.7 billion years ago, increased photosynthetic production of oxygen and higher net abundance could be safely sustained. Then, the atmosphere and surface ocean reservoirs became aerobic, with the deep ocean oxidizing. Banded iron formations could no longer be produced (except during extraordinary times of near-global glaciation) because iron dissolved in seawater was always unstable. The lack of iron meant that iron sulfides, which were an effective trap for sulfur, could no longer form, and deep ocean waters may have been sulfide rich. Redbed formations became more widespread.

17.7.4 Stage 4

Eventually, the deep ocean received enough flux of oxygen to become aerobic as well, ending the period of the "sulfide ocean" (Figure 17.4). The advent of oxygen respiration (aerobic metabolism) was initiated among living forms, and the number and vigor of photosynthesizing species increased. The new balance in oxygen production and loss was between photosynthesis and respiration/decay, with photochemistry, weathering, and volcanism now insignificant in their effect on oxygen levels. The balance was such as to permit a gradual increase in oxygen to the current abundance within the past billion years, with a handful of outstanding fluctuations upward and downward due to changes in burial rates of volcanism and organic matter, and glacial episodes.

17.8 Shield against ultraviolet radiation

The damaging short-wavelength ultraviolet (uv) photons from the Sun are today shielded by O_3 in Earth's stratospheric layer (Chapter 14). Ozone is produced photochemically from molecular oxygen (O_2) by absorption of uv photons.

Based on chemical models, to maintain an ozone shield requires 10^{-2} PAL or higher of oxygen. Clearly, then, an ozone shield was not available up to about 2 billion years ago. Because uv radiation is absorbed more effectively by water than is visible radiation, photosynthesizing organisms in the oceans could have been protected from uv radiation, even at shallow depths. Other atmospheric gases and aerosols, such as sulfur-bearing molecules, also might have afforded protection from some of the uv radiation, making life on land surfaces possible. Because these shields likely were not as effective as the current stratospheric ozone layer, organisms had to develop protection themselves; the common tendency of bacterial colonies to form mats, the evidence of which is the stromatolites, would have shielded such colonies from the uv flux. In spite of various survival strategies, incomplete shielding of continents and the ocean's surface from uv radiation probably restricted severely the number of

viable life-forms and viable habitats on the Archean and early Proterozoic Earth. It may also have been an aid to oxygen-producing photosynthetic organisms, which must have evolved to tolerate an environment in which oxygen-bearing free radicals were produced by uv photons acting on the atmosphere near the Earth's surface.

17.9 Onset of eukaryotic life

The dramatic rise in oxygen levels around 2 billion years before present resulted in two events that enabled a large increase in the forms and number of living organisms, and the ecological niches that they could occupy. These were (i) the enabling of aerobic respiration, which dramatically increased the energy that life could generate and use from the environment, and (ii) the development of an ozone shield.

All aerobic cells, be they prokaryotic or eukaryotic, contain enzymes that are required to detoxify the molecular fragments, or *radicals*, that contain oxygen. Without such enzymes, these free radicals would react with and destroy cellular structures. Anaerobes must avoid oxygen by existing in oxygen-poor environments or mounting defenses against oxygen similar to those of the aerobes. Even more intriguing is that, with just a few exceptions, oxygen is not used in the chemical pathways synthesizing proteins and other biological molecules – it is just used as an energy source. Perhaps it is simply too difficult a molecule to use, or perhaps this is yet another piece of evidence for the late onset of abundance O_2 in Earth's atmosphere.

Although some prokaryotes evolved to take advantage of oxygen and employ it in their metabolism, the advent of abundant oxygen in the atmosphere and ocean led to the successful spread and diversification of a new kind of cell. Around the 2-billion-year mark, in the mid-Proterozoic, fossil evidence of eukaryotes appears, in which cellular function is divided among individual areas (the organelles described in Chapter 12) separated by membranes and, in some cases, containing their own separate DNA and RNA. The cell's central genetic code is isolated in a nucleus, and organized into *chromosomes*; there is far more genetic material wrapped in the chromosomes than in the single strand of DNA contained in prokaryotes. (However, bacteria are far more genetically flexible than eukaryotes in that they readily pick up mobile packages of genes from other bacteria, allowing drastic changes in structure and function. Such package transfers to eukaryotes, the viruses, for example, almost always disrupt cell function.)

Essential to the workings of the eukaryotes in the current biosphere are the plastids (for example, the green chloroplasts) and the mitochondria, defined in Chapter 12. The plastids convert sunlight, carbon dioxide, and water into sugars. The mitochondria take alcohols and lactic acid – products of fermentation of food products that takes place in the cytoplasm of the cell – and conduct a set of chemical reactions involving oxygen and the fermentation products to create the enormous phosphate-bond storehouse of energy characteristic of aerobic metabolism.

The mitochondria and plastids are important also for providing a clue to the origin of the complex eukaryotes: both resemble bacteria. Mitochondria have their own DNA, messenger RNA, transfer RNA, and *ribosomes* (the sites of protein synthesis in prokaryotic and eukaryotic cells) within the mitochondrial membrane. The DNA floats within the mitochondria as strands, and is not bound in chromosomes. The ribosomes look like bacterial ribosomes, and are sensitive to the same antibiotics. Mitochondria divide at times different from the rest of the cell, by simple pinching and division, as do bacteria. Plastids resemble bacteria even more than do mitochondria in the appearance and arrangement of their internal structures.

The late biochemist Lynn Margulis proposed some years ago that the eukaryotic cell is the result of symbiotic (cooperative and dependant) relationships between bacteria of various types. Sometime in the past, presumably in the mid-Proterozoic as aerobic metabolism became possible, various symbiotic relationships between aerobic bacteria, cyanobacteria, and larger host bacteria created combined organisms that survived and prospered, eventually becoming fully internally dependent such that the resulting composite cells were the eukaryotes that we are made of today.

Although mitochondria and plastids cannot exist outside of their own cells, there are plenty of examples of symbiosis among bacteria, and between bacteria and eukaryotes, in both the natural world and in laboratory experiments conducted over the past few decades. Some eukaryotes are actually anaerobic, lacking mitochondria but, in some cases, containing organelles specialized for fermentation; examples include the protozoan *Giardia intestinalis*, responsible for severe diarrhea in humans. Some anaerobic eukaryotes exist in a tightly dependent relationship with other organisms, including bacteria, or actually harbor bacteria within their cells in a symbiotic relationship. Removal of the bacteria usually leads to death of the host eukaryote. In one case the symbiotic bacteria belong to a group generally thought to be good candidates for the ancestors of mitochondria. Laboratory experiments have successfully forced symbiosis between bacteria and amoebas that do not normally engage in such processes; by then selecting the amoebas that best accommodated the invaders, a colony of healthy amoebas was created. The bacteria lived off the amoebas and, curiously, the amoebas became dependent on the bacteria as well.

A further clue to the origin of eukaryotes lies in the predatory nature of some bacteria that will invade the cell walls of other bacteria. Although most such encounters eventually result in the death of the host, and hence of the invaders, in some cases the prey have evolved a tolerance for the predatory bacteria.

Margulis and others have proposed that several extant aerobic (and predatory) bacteria are descendants of bacteria that evolved into mitochondria. A large photosynthesizing bacterium, *prochloron*, with unusual plant-like properties and a taste for symbiosis in sea animals, is perhaps descended from a similar bacterium that infected certain cells and evolved into plastids. Similarly, the large host cellular mass of eukaryotes is echoed in the large bacterium, *thermoplasm*, that is modestly oxygen tolerant. Other candidates for the cell nucleus and additional eukaryotic cellular structures have been proposed.

The notion that complex plants and animals, including humans, are the result of symbiotic relationships between bacteria may be shocking to some, but it is increasingly accepted by biologists. The structural similarities between organelles and some bacteria, the symbiosis between anaerobic eukaryotes and bacteria, between different types of bacteria, and the tolerance

of some eukaryotes for forced laboratory symbiosis all suggest that such dependencies have arisen throughout the history of Earth.

Comparison of the genomes of eukaryotes and prokaryotes have led some molecular biologists to propose an earlier origin for eukaryotes, perhaps prior to 3 billion years ago. This interpretation of the data is controversial, and most biologists argue that eukaryotes most probably first appeared around 2 billion years ago. Nonetheless, one could imagine early bacterial experiments in symbiosis leading to the production of complex anaerobic cells that did not survive the rising tide of oxygen, perhaps because bacterial precursors to mitochondria were not available or could not be incorporated.

Regardless of the original appearance of eukaryotes, their success was in the utilization of atmospheric oxygen. As cyanobacterial photosynthesis polluted the atmosphere and shallow ocean with oxygen, some bacteria evolved to be oxygen tolerant and then oxygen dependent. The ancestors of mitochondria were efficient enough at using oxygen that their symbiotic relationship with host bacteria created an energy-efficient, adaptable cell that would become the basis for animals and fungi. Further symbiosis brought photosynthesis into the eukaryotic realm. Eukaryotes spread into a variety of ecological niches – some even adapted to survive in anaerobic environments, in contrast to those eukaryotes that lack mitochondria and were anaerobic from the start.

Interestingly, the advent of aerobic eukaryotes continued the trend toward increased morphological diversity and decreasing chemical variety that was discussed at the end of Chapter 13. Bacteria exhibit a wide range of metabolisms used to derive energy from the environment, including fermentation, sulfur metabolism, nitrogen consumption, and aerobic combustion of hydrogen. On the other hand, aerobic eukaryotes threw their lot in with the mitochondria, so that, in spite of the enormous diversity of shapes and types of multicelled animals and fungi, the power source is almost entirely aerobic respiration through the mitochondria. Most nucleated cells have essentially the same kind of metabolism. This metabolism, coupled with plant and bacterial photosynthesis to create oxygen, largely determines the atmospheric composition that we see today.

The advent of aerobic eukaryotes enabled predator–prey food chains to come into existence. Although predator–prey relationships exist among bacteria, the food chain is one level: predator bacteria invade prey bacteria, and the food chain ends there. Anaerobic organisms are not efficient enough at producing energy from fermentation and other mechanisms to create enough food for a multilevel food chain, and aerobic bacteria do not come in enough morphological varieties to create such a chain themselves. Eukaryotes have the high energy efficiency and morphological diversity to sustain the multilevel food chains that every student learns about in biology classes. In this regard, biologists T. Fenchel and B. J. Finlay point out that evolution toward large size and complexity is a tremendous evolutionary advantage – one can swallow smaller organisms or avoid (through size, speed, and smarts) being swallowed in turn. The high-efficiency oxygen metabolism and complex versatility of eukaryotes were prerequisites to innovations such as ourselves.

Following on the heels of symbiotic creation of aerobic eukaryotes, cooperative colonies of eukaryotic cells developed into the first multicellular organisms – plants and animals. By the end of the Proterozoic, a bit more than a half billion years ago, the O_2-rich, CO_2-poor atmosphere of Earth supported – and was sustained by – a wide variety of eukaryotic, multicellular species. This was perhaps the last step in the departure of Earth from the history of its neighboring planets; our atmosphere would never be the same. Nonetheless, many of the same external influences affected Earth as well as the other planets– variations in the Sun's brightness, shifts in orbits, occasional impacts of large objects to form craters. It is the response of Earth's atmosphere, oceans, and biological systems to such events that make up much of the story of the last half-billion years of Earth history – a story to which the remainder of this book is devoted.

Summary

The most profound change in the Earth's atmospheric and surface evolution came about when molecular oxygen levels increased greatly in the atmosphere. The present-day oxygen cycle is dominated by photosynthesis as the source, and respiration as the sink. These two processes are approximately in balance. A secondary source of oxygen – much less important, is photochemistry in the stratosphere, while weathering of rocks, volcanism, and plate-tectonic recycling of buried sediments serve to remove molecular oxygen, but at a much lower rate than respiration. Prior to the widespread development of photosynthesis, only photochemistry produced oxygen, and respiration did not exist; the other sinks kept oxygen at extremely low levels relative to today. The timing of the increase in oxygen is determined from the geologic record; minerals that would not be stable in an oxygen-rich environment such as uraninite and pyrite do not occur more recently than 2.7 billion years ago. Banded iron formations, a type of sediment that consists of alternating layers of iron- and silica-rich minerals, are relatively common between 2 and 3 billion years ago, but very rare prior to that period and thereafter. They seem to record marine environments in which oxygen levels fluctuated on various timescales, perhaps seasonal or longer. Evidently, as

photosynthesis became widespread, oxygen levels rose but not steadily since sinks of oxygen were still available, and the rate of photosynthesis varied seasonally or due to shifting climate. Thus different components of the Earth's surface – the atmosphere, shallow ocean, and deep ocean – became saturated in oxygen at different times. Eventually, the sinks of oxygen were saturated and the atmosphere and ocean became aerobic, although oxygen-poor environments still occur today. The great oxygenation of the Earth's ocean and atmosphere opened the way to the era of complex cells that undergo respiration, and after another billion years or more – to the complex animal and plant life with which we are familiar today. Alternatives to this story have been offered, including an early increase in oxygen, and possible changes in the mantle oxidation state as a major contributor to surface oxygen. But thanks to the fossil record we have little doubt that only late in our planet's history has the balance between photosynthesis and respiration been an intimate part of life's story.

Questions

1. How might the history of oxygen on Earth have been altered if the oceans were very shallow – only 100 meters in depth, for example?
2. How in turn might this have altered the evolution of life and the sustained habitability of Earth?
3. The story of oxygen is complicated by a chicken-and-egg problem: the kind of photosynthesis that produces oxygen – Photosystem-II – requires cells than can handle the presence of free radicals, which are the toxic byproduct of O_2 production. For such cells to have evolved to be oxygen tolerant in the first place, their precursors must have been exposed to molecular oxygen. Besides photochemistry in the atmosphere, what other sources might have contributed to the oxygen budget in localized places or times on the Earth?
4. Photosynthesis appears to be key to the production of oxygen. Yet on planets around M-dwarf stars, very little blue light is available. Investigate whether there are photosynthetic systems in life that can use longer wavelength, red light. Do these systems produce oxygen? What might be the evolution of the atmosphere be like on a planet orbiting a red dwarf star?

References

Anbar, A. D., Duan, Y., Lyons, T. W. *et al.* 2007. A whiff of oxygen before the Great Oxidation Event? *Science* **317**, 1903–6.

Canfield, D. E. 2004. Evolution of the Earth surface sulfide reservoir. *American Journal of Science* **304**, 839–61.

Cloud, P. 1988. *Oasis in Space: Earth History from the Beginning*. W. W. Norton, New York.

Fenchel, T. and Finlay, B. J. 1994. The evolution of life without oxygen. *American Scientist* **82**(1), 22–9.

Hoffman, P. F., Kaufman, A. J., Halverson, G. P., and Schrag, D. P. 1998. A neoproterozoic Snowball Earth. *Science* **281**, 1342–6.

Holland, H. D. 2009. Why the atmosphere became oxygenated: a proposal. *Geochimica Cosmochimica Acta* **73**, 5241–55.

Kasting, J. F. 1991. Box models for the evolution of atmospheric oxygen: an update. *Paleogeography, Paleoclimatology, Paleoecology* **97**, 125–31.

Kasting, J. F., Howard, M. T., Wallmann, K. *et al.* 2006. Paleoclimates, ocean depth, and the oxygen isotopic composition of seawater. *Earth and Planetary Science Letters* **252**, 82–93.

Kirschvink, J. L. and Kopp, R. E. 2008 Paleoproterozic icehouses and the evolution of oxygen mediating enzymes: the case for a late origin of photosystem-II. *Philosophical Transactions of the Royal Society of London, Series B* **363**, 2755–65.

Knoll, A. H. 2011. The multiple origins of complex multicellularity. *Annual Reviews Earth and Planetary Science* **39**, 217–39.

Margulis, L. and Sagan, D. 1986. *Microcosmos: Four Billion Years of Microbial Evolution*. Summit Books, New York.

Pavlov, A. A. and Kasting, J. F. 2002. Mass-independent fractionation of sulfur isotopes in Archean sediments: strong evidence for an anoxic Archean atmosphere. *Astrobiology* **2**, 27–41.

Press, F. and Siever, R. 1978. *Earth*. W. H. Freeman and Company, San Francisco.

Rasmussen, B., Fletcher, I. R., Brocks, J. J., Kilburn, M. R. 2008. Reassessing the first appearance of eukaryotes and cyanobacteria. *Nature* **455**, 1101–4.

Rosing, M. T. and Frei, R. 2003. U-rich Archaean sea-floor sediments from Greenland – indications of >3700 Ma oxygenic photosynthesis. *Earth and Planetary Sciences Letters* **6907**, 1–8.

18

The Phanerozoic: flowering and extinction of complex life

Introduction

The Phanerozoic eon is a major division in the fossil record that dates radioisotopically at a bit younger than 600 million years before present. Its geologic marker is the appearance of numerous complex multicellular organisms in the fossil record. This eon has no counterpart on any other planet, even if Mars harbored simple life-forms within the first billion years of its history. On Phanerozoic Earth, life began to occupy just about every conceivable niche on land, sea, and air. Geologically, Earth was more or less modern in form as the eon opened: the total continental mass was comparable to that today, modern-style plate tectonics were operating, and oxygen levels in the atmosphere were approaching present-day values.

The Phanerozoic eon is divided into eras, eras into periods, and periods into epochs. The boundaries between most of the periods are defined by extinction episodes in which a number (sometimes very large) of species disappear and are replaced in the sedimentary fossil record above that point by new species. Although the resulting story of complex multicellular organisms is too large to tell in detail in this book, some of the highlights are shown in Figure 18.1.

The presence of multicellular organisms per se was not new. Multicellular bacterial colonies had existed since the Archean; multicellular algae (for example, green seaweed) made their appearance shortly after the first unicellular eukaryotes in the fossil record. In each of these, and many other cases, there is little or no specialization among cells, and only limited communication. The Phanerozoic biological revolution was about organisms composed of cells, the forms of which were altered to conduct specific functions, and which were wholly dependent on one another. Animals are the extreme expression of this intricate symbiosis; plants exhibit this to a lesser extent.

What precipitated the rapid flowering of life into the diverse and complex multicellular entities of the modern era remains a mystery. Biologists, particularly those whose specialty is the understanding of the genetic code, tend to look for answers in the nuclei of eukaryotic cells, to see whether the vastly increased complexity of the genes precipitated this new way of living. At the heart of the mystery, though, is why it took a billion years after the appearance of eukaryotes for such an innovation to become widespread. Some recently found fossils, collected in China, suggest that multicellular plants existed at the 1.7-billion-year mark, but the rapid proliferation of complex multicellular life-forms did not occur for another billion years – 20% of Earth's history.

Geologists and planetary scientists tend to look for external causes behind the great flowering of complex life at the dawn of the Phanerozoic. Indeed, over the past several decades, the importance of environmental changes as a stimulus for major evolutionary changes and extinctions has been recognized. Even more recently, the possibility that impacts of asteroidal or cometary fragments could be the cause of episodes of large-scale extinctions has gained respectability. As environments change or whole ecosystems are wiped out, new forms of life develop from old to occupy the now-empty niches.

In fact, and as is realized by biologists and physical scientists alike, the flowering of whole new phyla of life-forms and their elaboration through the Phanerozoic require both internal genetic change and the external stimuli to make them happen. Coupling these leads to a view of evolution that is much more complex than Darwin's eternal set of battles for survival of the fittest member of a species played out against the background of a gradually changing Earth. The fossil record seems to demand a more complicated view.

In this chapter we examine the internal and external mechanisms behind the evolution of species. We focus on two major biological events in Earth's history: the flowering of multicellular life at the start of the Phanerozoic and the extinction of many families of species at the close of the Cretaceous period some 65 million years ago. These are by no means the only watershed moments in the biological evolution of the past 15% of Earth's history, but they are illustrative examples of two different kinds of evolutionary events – one perhaps more internal than external, and the other quite the opposite.

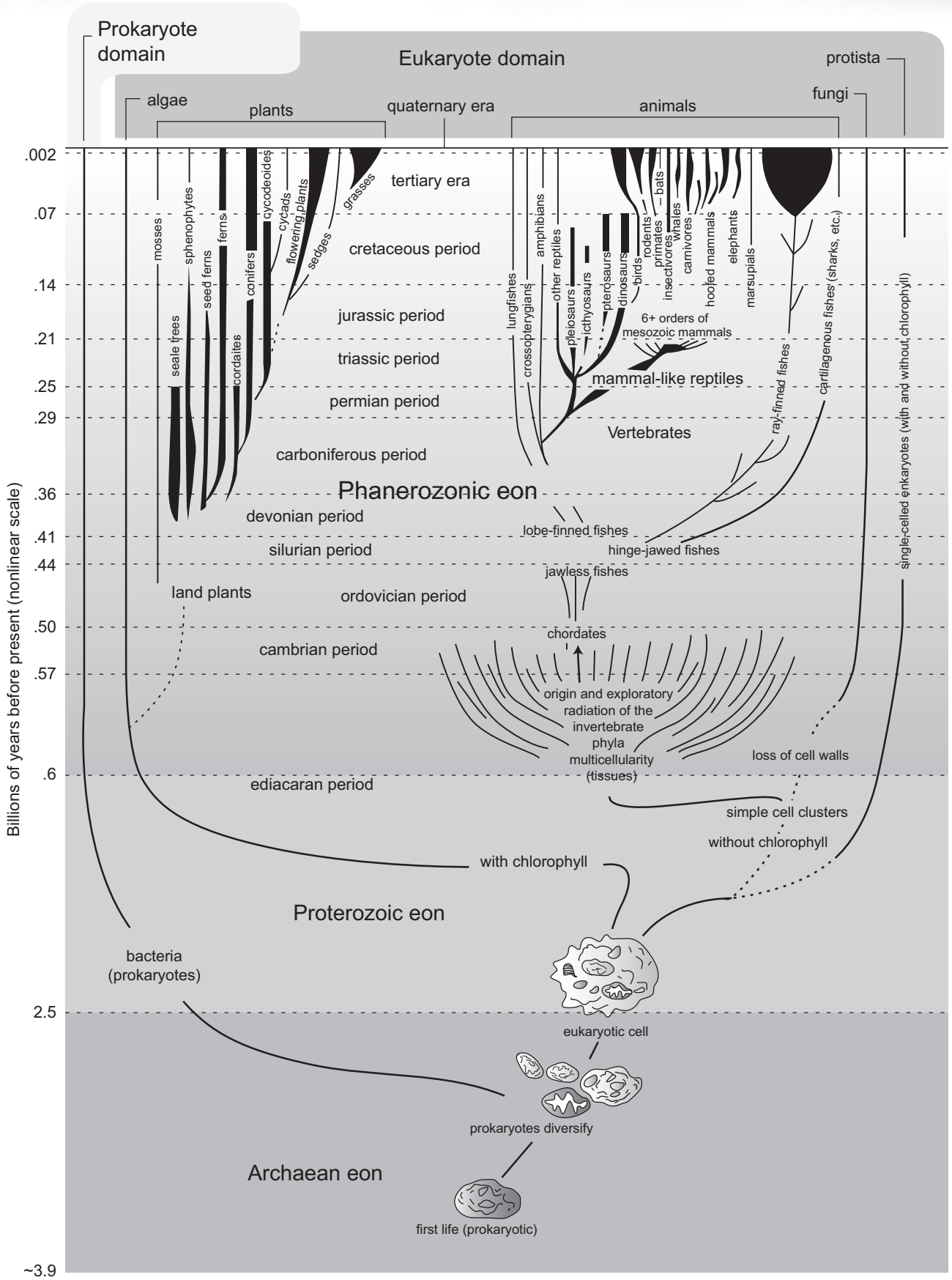


Figure 18.1 Major biological events in the Phanerozoic set against the backdrop of earlier eons. The chart is divided into prokaryotes and eukaryotes, and among the eukaryotes into the major kingdoms. Eukaryotic lines of descent shown at earliest times are phyla, then in later times some (but not all) major classes and orders of vertebrates and plants are shown. Of note is the very large number of different animal phyla in the early Phanerozoic, some of which are no longer represented by species. An attempt to show the range of diversity of forms in the plants and vertebrates is given by the thickness of the lines. Particularly uncertain aspects of the histories are indicated by dashed lines. Redrawn and modified from Cloud (1988).

18.1 Evolution

Much controversy still surrounds the concept of the evolution of species. The controversy is not so much scientific as it is political, centered on the question of the special and simultaneous creation of all extant species as a literal interpretation of the Judeo–Christian bible would require. It is not the place of this book to argue the merits of this point of view on spiritual grounds, but the following are offered in support of the notion that species evolve with time:

1. The fossil record shows a wide range of extinct life-forms that bear an increasing relationship to current living organisms in progressively younger sediments.
2. The source of the instructions by which the form and function of organisms are defined, DNA and RNA, was identified almost a half-century ago. Since that time, the ability to manipulate and transfer genetic code to effect a change in form and function of organisms has been repeatedly demonstrated; this *recombinant DNA* technique is now routinely used to make agricultural and pharmaceutical products.
3. Natural selection and consequent evolution have been directly observed for a few species with short reproductive cycles that have experienced sudden changes in their environments.

Broadly defined, *evolution* is the formation of new species from old. Viewed in a slightly more specific way, it is the change of form and function of living organisms sufficiently profound to create a new, self-contained breeding population. The two definitions are tied through the concept of *species*, a taxonomic label whose precise characterization has proved difficult. We recognize intuitively the human species – all humans alive on Earth today are the same species and can produce offspring with each other. Our ability to distinguish between members of different species versus more general taxonomic orders (see section 18.3) decreases as the relation to us becomes more distant. A chimpanzee is a chimpanzee – but the bonobo "chimps," it turns out, are a different species. It is much harder to decide what constitutes a species when staring at different kinds of fungi – unless one is a professional biologist.

The best and simplest definition of species, at least for animals, is that it is a cohort of living organisms that can produce viable and fertile offspring that interact sexually and hence are isolated from other organisms in the reproductive sense. The qualification "fertile" is required because different species – tigers and lions, or horses and donkeys – can produce common offspring that are, however, sterile. This definition does not work as well for many kinds of sexually reproducing plants; and among forms of life that do not reproduce sexually, or do so in the "bacterial fashion" through frequent exchange of DNA fragments, the concept of species is almost irrelevant.

Evolution would not be possible if the reproduction of the genetic code were error free from generation to generation. However, alteration of the DNA code, or *mutation*, occurs by random copying errors and by errors induced by ultraviolet light (particularly at wavelengths of 2,600–2,800 Angstroms) and impacts of cosmic rays (high-energy protons streaming through Earth's atmosphere) on the DNA molecule.

Mutations induced in the sexual cells – eggs and sperm – may result in a change of form or function of the offspring. Some mutations are harmful, many are fatal, but a large number – those that alter certain codons such that no new amino acid is specified for that position – are neutral. However, others may confer some advantage to the offspring and to future generations in which the new genetic code is perpetuated. This brings into the picture the second key requirement for Darwinian evolution, which is *natural selection*: the nature of the environment (including interaction with other organisms) exerts pressures on individuals and species which will tend to amplify the effects of some mutations. In some cases, particularly when the environment undergoes a change or small groups are isolated by environmental effects, new species can arise from old ones.

Although the genetic code is voluminous, recent discoveries of key gene sequences, or "trigger" genes, suggest that some single mutations may have dramatic effects. Trigger genes are those that control the activation of large sets of genes, which in the aggregate control a major structural or functional characteristic of a species. In fruit flies, for example, the activation of a single or small number of trigger genes means the difference between development of a leg versus an antenna.

18.1.1 Classical Darwinian, model of evolution

Given a mechanism for change in species, how does evolution actually work? The traditional mechanism, still taught in high school biology classes, was first clearly articulated by Darwin, and is popularly called "survival of the fittest." In this view, individuals compete within a more-or-less stable or slowly changing environment; individuals with the characteristics that make them most competitive survive to produce multiple offspring, often with multiple reproductive partners, offspring who carry those characteristics.

The implications of this view for evolution are that the transformation of ancestral populations into a new species is slow and fairly constant in time, and a large fraction of the population, over most of its geographic range, is transformed. Evolution is a gradual process, in this view, with two important implications for the fossil record:

1. An ideal sedimentary (fossil) record of the origin of new species will contain a sequence of forms intermediate between the ancestral and descendant species.
2. Breaks or gaps in the forms intermediate between the old and the new species are caused by the intrinsic incompleteness or imperfections in the sedimentary record.

Examination of the fossil record shows few or no intermediate forms between the new species. One objective interpretation is that species do not evolve, but arise out of whole cloth. The other interpretation is based on mechanisms of fossilization (Chapter 8), namely that fossilization is so rare and the record so imperfect that the chance of catching a species in the act of changing is vanishingly small. This latter argument has been

used by geologists for roughly a century, and it is not unreasonable: for a body form to be well preserved in the sedimentary record, it must be allowed to remain intact against bacterial decay and physical damage. Once fossilized, the sediments themselves must remain relatively unaltered through enormous spans of time. As noted in Chapter 8, only a very small fraction of the world's organisms have had the "privilege" of becoming fossils.

The lack of transitional forms is a problem in spite of the difficulty of fossilization, and it provided fertile ground for antievolutionists to argue that the whole concept of natural changes in and formation of new species was wrong. Even Charles Darwin regarded the fossil record as an embarrassment for his theory. But the problem was rationalized away and new generations of paleontologists (those who study fossils), raised on the textbooks and canons of their mentors, carried on the tradition of arguing that the absence of transitional forms reflected the imperfections of the geologic record. As more and more sedimentary layers were analyzed and revealed more detail in the progression of life-forms – without solving the transition problem – paleontology reached something of a crisis by the 1960s.

Science is a self-correcting process, and flawed hypotheses find themselves defeated by the sword of new data. But often a whole picture or paradigm of the way things work, maintained over decades or more of work in a particular field, finally succumbs to a mountain of evidence. In this case, a wholly new paradigm usually comes to the fore, and often we view this in retrospect as a revolution in science. The development of quantum mechanics to supplant classical physics was one such revolution. The development of a new model for the way evolution works arguably could be called a minirevolution, one that introduced the concept of *punctuated evolution*.

18.1.2 Punctuated equilibrium approach to evolution

On the island of Bermuda lies an excellent fossil record of the history of a particular snail species, over the last 300,000 years of Earth history (a period we cover beginning in Chapter 20). The sedimentary record is well preserved, a large number of fossils occur in a fairly small area, and one subspecies is currently in existence. (A subspecies, even more difficult to define than a species, is a population with distinctive characteristics but still able to produce fertile offspring with other members of the larger species. That is, different subspecies of the same species interact genetically with each other, whereas different species do not. Breeds of dogs, or horses, constitute examples of subspecies.)

Harvard paleontologist Stephen Jay Gould studied this high-fidelity fossil record over a quarter of a century ago and came to a surprising conclusion: he could understand the formation and disappearance of new subspecies, not in terms of gradual evolution, but in terms of isolation and subsequent rapid changes in the characteristics of a given population of snails. Characteristics studied included variations in the color banding of the shell, in the general form of the shell's spire and lip, and in its thickness.

The details of Gould's analysis of the Bermuda fossil record are not repeated here, but the interested reader is referred to the article by Eldredge and Gould (1972). Figure 18.2 illustrates the relationships in time among the various subspecies of snails, based on the sedimentary sequence. In terms of timescales, the production of new subspecies was very rapid and occurred in populations at the edges of the main species itself. New subspecies of the snails appeared on the geographic periphery of a population of established snails with similar body characteristics. Body characteristics of a western subspecies did not appear on the periphery of the eastern subspecies, and vice versa. Eventually, the original subspecies became extinct, leaving the newer ones to continue in the fossil record.

Other examples are cited by Eldredge and Gould in the fossil record, wherein new species are seen appearing suddenly in the fossil record at the geographic periphery of existing species, with body types clearly most closely related to that proximate population. At issue is how such new species could have evolved out of the old in a relatively short time span. Eldredge and Gould suggest that the new population, being on the periphery of the old, becomes isolated geographically, perhaps through migration, climate change, sea level rise or fall, or other factors. The members of this population then interbreed among themselves, so that particular characteristics (including recessive ones) of that genetic pool are distributed and enhanced through that population. In cases where new features are neutral or advantageous with respect to survival, the isolated population as a new subspecies or species might well outlast and replace the old.

That such isolated populations occur is amply illustrated by the cheetah, which today is remarkable in the enormous genetic similarity among different members of the species. Typical species, including humans, exhibit much greater genetic variety among individuals than does the cheetah. Sometime in the past, most of the cheetah population was destroyed by environmental changes, or perhaps by human hunting. Only a small, isolated, interbreeding subgroup remained. Inherent in this genetic identity lies the danger of a single disease wiping out the entire population, but such identity also implies that the particular special characteristics of this group are now part of what defines "a cheetah."

The bottom line of the punctuated equilibrium evolutionary model is that the fossil record accurately but incompletely reflects the history of the species contained within it. Species characteristics are stable over long periods of time, and then isolated populations of a given species may change rapidly as they breed among their limited group. The potential for change lies within the genetic code, expressed as an existing characteristic that gets amplified or more prevalent among certain members through interbreeding, or a random DNA mutation leading to an advantageous new characteristic or function. Evolution does *not*, in this view, proceed by gradual competition and elimination of the less fit members of a large breeding population; the action occurs among the few, in isolation, perhaps at times of great environmental stress.

Eldredge and Gould's idea was and still is controversial. It clearly explains some aspects of the fossil record, particularly its inability to show gradual morphological (shape) changes between one species and a clearly related one in younger sediments. But some changes seem so abrupt that punctuated equilibria cannot be the entire explanation for the nature of the fossil record. As sedimentary layers of rock are subjected to new tectonic forces in a region, they may be overturned, partially altered,

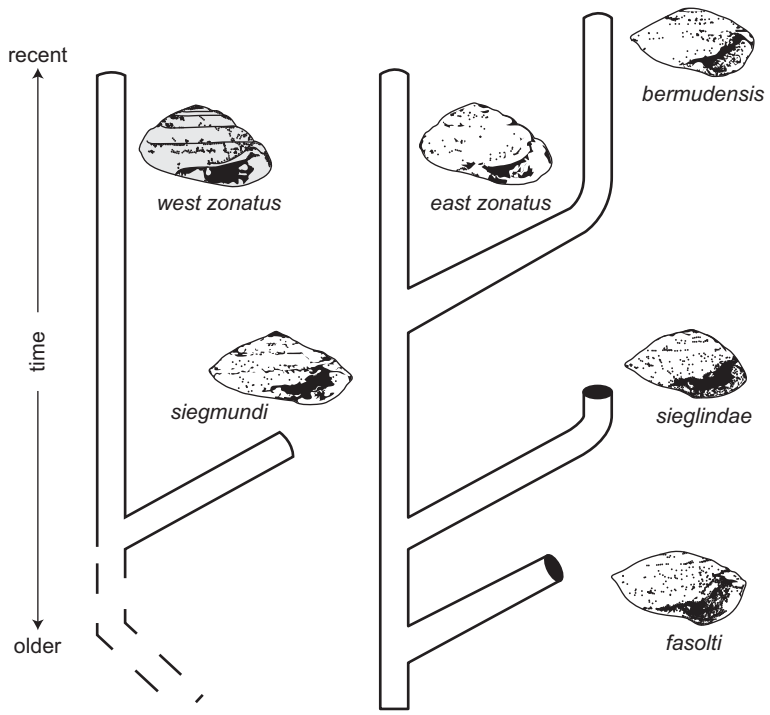


Figure 18.2 Results of Gould's analysis of populations of snails in Bermuda. The vertical axis is time, increasing upward. The branches are used to show the named subspecies branching off from the main species. Such branchings occurred on short timescales compared to that covered by the entire sequence. Although not indicated on the graph, these branchings occurred in geographically isolated populations. The two main species, *west zonatus* and *east zonatus*, presumably diverged from a common ancestor at a time well before that represented on the graph. Redrawn from Eldredge and Gould (1972).

or destroyed, taking key pieces of the fossil record with them. The vast majority of organisms that have lived on Earth did not have the honor of becoming fossils; their body parts were consumed by other organisms, including bacteria, to a complete or nearly complete extent. Finally, major disasters in Earth's history, one of which we document below, killed off species (often large numbers of species) very quickly, creating a break in the paleontological record.

For these reasons, the abrupt transitions in the fossil record must be a combination of imperfect preservation, breaks due to geologic processes and major extinction events, and the tendency of species to undergo rapid changes in isolated populations. Punctuated equilibria may well be the whole story in evolution, but it is only part of the story in explaining the appearance of the fossil record itself.

A lingering question that must be asked is whether the nature of the genetic code allows for rapid and ultimately profound changes in species characteristics. At one level, the punctuated equilibrium model doesn't make any demands on the genetic code: it simply requires that a small breeding population tend to enhance the particular special characteristics of group members, and we know this to be the case from breeding dogs or royalty. However, the genetic code must have within it the capacity to enable the remarkable development of intricate structures such as eyes, heart, brain, and hands, regardless of how evolution proceeds (that is, by punctuated equilibrium or otherwise).

The discovery of classes of genes that trigger, or turn on and off, whole groups of other genes, leading to drastic differences in form and function, provides a mechanism for dramatic changes in the morphology and functionality of eukaryotic cells and multicelled organisms. These and other features of the genetic code seem to be capable of priming organisms for significant changes in their appearance and complexity. Over the enormous spans of geologic time, specialized cells and organs bootstrapped new functions out of old. But not in a gradual fashion; punctuated equilibrium implies long periods of stability in a given species followed by rapid, and perhaps dramatic, change. What may happen during those long time spans of stability is that mutations that are individually neutral in their effect accumulate in organisms and are passed down over generations, ultimately until a mutation in a trigger gene causes a large number of the existing mutations to finally have an effect on the organism.

Whether triggering genes or other components of the complex genetic code provide the key to the enormous variations seen in biology remains an open question, because we exist still in the youth of the genetic revolution. But that such a revolution exists at all is one of the wonders of twenty-first century science. As recently as the nineteenth century, the human egg was regarded as bearing homunculi: fully formed but miniature human beings awaiting the sperm to make them grow. The discovery of DNA followed by modern techniques of mapping the genetic code and manipulating genes have revealed a richness of malleable information at the heart of every cell that provides the enormous potential both for stability and change. Eukaryotic cells contain far more genetic material than is actually used to code for body type and function; the purpose of the rest remains unclear.

The rapid production of new forms and hence new species is not the purview of sexual reproduction alone. Many creatures clone themselves; some (like aphids) do so at particular times of the year and then later conduct sexual reproduction. The acquisition of extra chromosome pairs, crucial to development of commercial cotton, wheat, and other crops, also may play a role in development of new species. The splitting of chromosomes at their tie points occurs occasionally, and there is evidence in chromosome maps that some mammalian species are derived from others by this process (dogs from wolves, for example). Although the usual result of extra chromosomes is harmful (Down syndrome, for example), the genetic evidence tells us that there are occasional cases where something advantageous happens.

We need not understand the detailed working of the genetic code to understand that species are, in fact, complex dynamical systems. They draw energy through their cells, conduct regulated chemical processes to sustain and reproduce those cells, and they produce copies of themselves. Complex dynamical systems have the property of remaining in one sort of stable state or mode for some time and then suddenly, through a relatively small change in the external conditions, undergoing drastic change to a different state. Species are like that: they exist in stable form for an enormity of generations (tens of thousands to millions) and then, perhaps stimulated by a changing environment, an isolated fringe of that species will shift in relatively few generations to a different form or mode of operation. Nothing supernatural need happen: just the availability and activation of a complex but malleable sequence of genes with a range of regulatory and control functions.

In Chapter 20, we examine the evidence for such rapid changes in the history of the precursors of modern humans, a story that is rich in long-lived hominid species punctuated by rapid changes. But long before those events came an extraordinary change in the nature of eukaryotic organisms, near the beginning of the Phanerozoic. It is this, the Cambrian revolution, that provides the most dramatic example of an explosion of novel body types in new species, one that defined the body plans for much of what was to follow in life's history.

18.1.3 Other models for evolution

Even if punctuated evolution is a good model for how new species arise from old, there are other evolutionary mechanisms. In Chapter 17 we were introduced to the notion of symbiosis leading to the complete dependency of one organism on another and, eventually, development of a single organism.

18.2 Ediacaran–Cambrian revolution

The time periods immediately preceding and following the start of the Phanerozoic eon, the Ediacaran (formerly called the "Vendian") and Cambrian periods, represent a biological fomenting unprecedented before and since that time. The fossil record of this time is seen in sediments that were originally seafloor environments and were later pushed onto the continents by plate tectonic processes to become part of the melange of preserved sediments. Most famous among these Cambrian sediments is the Burgess Shale, layers thrust upward in geologically recent times to form part of the British Columbian Cascade Mountains. However, many other, originally marine sediments from this time are found around the world. They all tell the same story of dramatic biological change. To appreciate the change that took place requires understanding a little of the way biologists classify organisms, a digression we turn to before examining the fossil record.

18.2.1 Taxonomy and phylogeny for the restless

A necessary evil is at least a cursory acquaintance with the system for classifying life, established in its earliest form some 250 years ago. This *taxonomic* system classifies species according to morphological (form-based) relationships into progressively higher and broader categories, ending with the major eukaryotic *kingdoms* of algae, plants, animals, fungi, and single-celled protists. (A higher level, *domain*, separates the two prokaryotic types from each other and from eukaryotes.) The order of the major taxonomic categories, from most to least specific, are species, genus, family, order, class, phylum, kingdom, domain.

A eukaryotic life-form can be identified in this scheme by its genus (capitalized) followed by the specific name – hence the domestic dog is *Canis familiaris*. The mountain coyote is *Canis latrans*; the grey wolf, *Canis lupus*. These and all other members of the genus *Canis* belong to the family Canidae, or true canines, which lies within the order Carnivora – flesh-eating mammals including cats, racoons, bears, hyaena, and many others. The class Mammalia is broader still, embracing the Carnivora, kangaroos, rats, aardvarks, dolphins, whales, humans, and others. The principal commonality is that all are vertebrates and all female mammals secrete milk for feeding their young. Mammals are part of the phylum Chordata, meaning animals with internal skeletons (hence excluding insects, crabs, etc.), central nervous systems and, with a few exceptions, possessing a vertebrate-type backbone and skeletal arrangement. (Strictly speaking, vertebrates are a subphylum of Chordata.) Finally, chordates are members of the animal kingdom, a multicelled eukaryotic organism, nonphotosynthesizing (hence not plant or algae), with distinct cell walls and a single nucleus per cell (hence not fungi).

The organization of creatures into this framework is an enormous undertaking, one that has been at the core of botany and other types of biological field endeavors. Although form was the original criterion, allowing extinct creatures to be classified on the basis of fossil morphology, sequencing of the genetic code to determine species relationships at the blueprint level is now a standard technique used on living species. Such an exercise might seem to some the ultimate in butterfly collecting, but it has revealed over decades of work the pattern of biological forms and has allowed inferences as to the genesis of particular species from each other. In the animal kingdom, 30 phyla exist, comprising perhaps 10 million species.

Phylogeny is the study of the connections between groups of organisms by understanding how species are related to each other by descent in time. The relationships among organisms are understood through *cladograms*, which can also be called *phylogenetic trees* – the species equivalent of human family

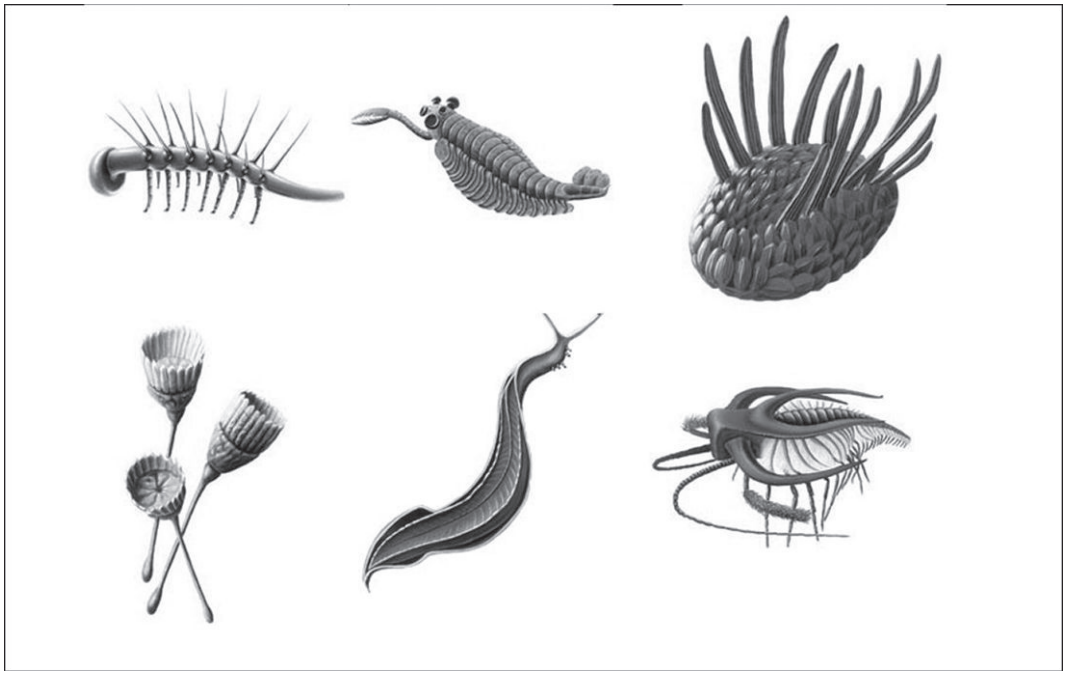


Figure 18.3 Some denizens of the Cambrian. Top row from left to right: hallucigenia, opabinia, and wiwaxia. Bottom row, left to right: dinomischus, pikaia, and marrella. Images with kind permission of The Burgess Shale Geoscience Foundation, Field, BC.

trees. A summary phylogenetic tree for all of life was shown in Figure 12.10 in Chapter 12.

18.2.2 Establishment of the basic plans

To appreciate the revolution at the start of the Phanerozoic requires recognizing that the taxonomic level *phylum* refers to basic, fundamental body plans. We instinctively think of insects and spiders (members of the phylum Arthropoda) as wholly different from us vertebrates because their body plan is based around a fundamentally different architecture – that of the exoskeleton. What is remarkable about the start of the Phanerozoic is that, within about 10 million years, all but perhaps one of the animal phyla in existence today appears in the fossil record. (This revolution is restricted to animals – higher plants began to appear in abundance as land-colonizing descendants of algae roughly 460 million years ago, after the Cambrian.) Included in the Cambrian shales is a fossil animal no more than two inches (5 cm) long with a clearly defined spinal rod, the 525-million-year-old ancestor of the Chordata – the mother of us vertebrates.

Understanding of the importance of the Cambrian explosion has grown, and especially in recent years. Examination of the Burgess Shale early in the twentieth century revealed the appearance of many invertebrate phyla, but only in the past 30 years has re-examination of the shales revealed the forms of many soft-bodied animals previously missed. In the past decade, examples of chordate ancestors have been found, and better radioisotopic dating, using volcanic ash with uranium-bearing minerals, has shrunk the whole revolution down in time to a narrow 10-million-year window.

Not all phyla made it. Roughly a third again as many extinct phyla are seen in the Cambrian record as exist today. Moreover, each phylum is represented by far fewer species than are contained in extant phyla today. Clearly, the remarkable radiation of body forms involved large leaps in structures, without the more minor elaborations and variations at the family-through-species levels yet to come.

Equally important in this biological revolution is what didn't happen later: at no other time in the subsequent history of animal life did such abundant and new innovations occur. Once formed, the many phyla of the Cambrian became all there is. When we look at the remarkable diversity of mammals, reptiles, and birds, for example, we see that all have a remarkably similar body plan, on which small changes in form and function are elaborated. A human-sized, twelve-legged, exoskeletal merry-go-round with an extendable mouth at the bottom and a crop of bushy hair at the top, foraging across the grasslands, is not in the cards because the basics of such a body plan did not appear in the Cambrian. Nonetheless, many of the Cambrian pioneers of the dozens of new phyla were monstrous in themselves, such as those seen in Figure 18.3.

18.2.3 Clues from the Ediacaran

What happened in the Cambrian need not invoke invention out of whole cloth. Recent careful work on fossil sediments from the period immediately preceding the Cambrian – the Ediacaran – has revealed increasing numbers of interesting, organized, multicellular animals. These sea creatures, existing over a time of several tens of millions of years, look very much like palm fronds arranged in linear, radial, or bilateral schemes. Some bear enough resemblance to worms, seaweed, and jellyfish to be argued as their ancestors, but the jury is still out. Are the Ediacaran creatures another failed experiment, on the eve of

the Cambrian, in multicelled organized animals? Or are they the seedcorn from which the Cambrian phyla exploded, a kind of biological bootstrap that allowed modern body plans to be achieved?

18.2.4 Causes of the Ediacaran–Cambrian revolution

Whether the Ediacaran biota were related to those in the Cambrian is secondary to the important innovation that was taking place: in a short span of geological time, complex and mobile creatures, organized of assemblages of eukaryotic cells, were appearing on a number of different body plans. Many of the innovations were relatively simple, animal hard parts being little different than the calcareous wastes or shells of more primitive unicellular creatures. More dramatic was the differentiation of cells into interdependent entities with distinct forms and functions. No long, gradual set of changes from multicelled, undifferentiated colonies to animals is seen – the eukaryotic menagerie remains relatively uninventive for at least a billion years prior to the Ediacaran – the "boring billion," as some who study Earth history have called it.

A number of possible factors, working together or separately, may have contributed to the Ediacaran–Cambrian revolution:

1. *Effects of near-global glaciations.* No less than three major glacial episodes occurred between 730 million and 570 million years ago. Ice may have covered much of the Earth, though the extent remains debated. The most extreme glacial episodes, in which ice may have extended nearly all the way to the equator, would have had a profound effect on climate. The ice separated the ocean from the atmosphere, reducing or eliminating the ocean's ability to absorb heat and buffer temperature differences on a global scale. The Earth's surface would have experienced much more dramatic day–night temperature extremes than it does today, and potentially more extreme seasonal variations from summer to winter – though ice cover limited summer temperatures. In some ways, a snowball Earth climate – bereft of the control of the oceans – might have been a bit like Mars, but Mars with a thick atmosphere. Such temperature swings and extensive ice cover would have limited the availability of suitable ecosystems for the primitive microbes then present, leading to massive extinctions. Such extinction events are difficult to see in the fossil record when the sum total of life on Earth is microbial; it is recorded instead perhaps in the fluctuating carbon isotopic ratios. This mechanism does not explain why animals are not seen in the geologic record two and one billion years ago – prior to the major glacial episodes and after the first occurrence of eukaryotes. But it may provide an explanation for the particular point in geologic history when animals did suddenly appear and survive: the end of the last deep ice age demarcates the appearance of primitive animal forms in the Ediacaran, a kind of early genomic springtime before the full bursting forth of animal life in the Cambrian.
2. *Sulfide ocean, oxygen levels, and carbon burial.* Macroscopic animal forms cannot exist in low oxygen level environments, even those that might sustain single-celled eukaryotes, because the energy cost of delivering sufficient oxygen to cells in the interior of the organism is too high. The sulfur-rich ocean between 0.6 billion and 2 billion years ago discussed in Chapter 17 could have exacerbated low oxygen levels in the Earth's oceans, allowing for eukaryotic cells but not for multicellular life. By 600 million years ago, oxygen levels were not far from the present values and were increasing. A decrease in the abundance of biologically preferred carbon isotope ^{12}C relative to the rarer and less-palatable ^{13}C, suggests that just prior to the Cambrian a large amount of carbon was being removed from continents and shallow oceans and deposited on the deep ocean floor. Removal of large amounts of carbon from the biologically active shallow marine environments could have boosted the levels of molecular oxygen, which otherwise would have combined with the decaying carbon to make carbon dioxide.
3. *Genetic complexity.* To take advantage of a new kind of environment requires that the genetic mechanisms be sufficiently complex to allow drastic changes in form and function. Eukaryotic cells have much more genetic material than prokaryotes, and in fact only a very small portion of it is actively used to control protein production and other functions. The latency present in underutilized genetic information is a potentially powerful force for change. If multicelled, complex animals could not have survived below certain atmospheric oxygen thresholds, that would not have prevented repeated experiments in this direction, all of which failed before becoming abundant enough to show up in the fossil record. When the oxygen levels rose above a requisite threshold, the next set of organizational attempts worked, and animals began to multiply and diversify.

   Some biologists have argued that the threshold trigger was not oxygen but genetics itself. More primitive multicelled animals have fewer regulatory, or trigger genes, than do more complex forms. Presumably there is a threshold number to achieve any sort of multicelled organization at all. In this view, the Ediacaran–Cambrian opened with a chance production of additional regulatory genes, presumably co-opted from the existing genetic codes available in eukaryotic cells. Possibly the Ediacaran innovation was oxygen related, followed quickly and coincidentally by the innovation of more trigger genes and consequent explosion of body types. Or perhaps the trigger genes were available first, followed by the rapid rise in oxygen.
4. *Absence of predators.* Another contributing factor to the rapid appearance of different body types was the absence of other creatures in the same ecological niches and, specifically, the absence of predation. Simpler unicellular or colonial eukaryotes could not take advantage of the environment in the way that complex multicellular animals could; hence, once the latter appeared, they had no competition. Many different experimental body plans would all have flourished in this environment, until competition became intense enough, because of crowding, that predation took place. Some paleontologists see evidence for predation beginning at the end of the Ediacaran.
5. *Artifact of the geologic record.* Until recently it was thought that the sudden appearance of Cambrian shelled organisms was an artifact of their descent from earlier, soft-bodied creatures. More detailed analysis of sediments revealed large

numbers of soft-bodied imprints in both the Cambrian and the Ediacaran, indicating that the sudden flowering of life was not an artifact of the data record. Nonetheless, one cannot discount entirely the possibility of future discoveries of animals in the fossil record *prior* to the Ediacaran.

Regardless of the specific causes, the Ediacaran–Cambrian revolution seems, in retrospect, inevitable. At issue is only the timing. It is quite possible, and intriguing to consider, that over the prior billion years similar experiments did happen. Absent enough oxygen, these creatures were ill fated; alternatively, they might have flourished in environments temporarily rich in oxygen. As oxygen levels dropped again, such tentative experiments disappeared, never abundant enough or sustained enough to show up in fossils except as hints here and there. Understood as a complex, self-organizing system (Chapter 13), life's giant step across the multicelled-body threshold is an inevitable behavior of a dynamical system. Change the conditions and the system changes its mode of operating. Provide plenty of energy flow and more complexity and new examples of organization will appear spontaneously.

18.2.5 Why has it not happened again?

More difficult than why is "why not again." It is possible that Cambrian and post-Cambrian levels of biological complexity fully exploit the genetic capabilities of our cells, at least within the current environment. Little has changed on Earth in the past half-billion years. Oxygen levels have stabilized, plate tectonics moves continents around, and those as well as orbital cycles cause the climate of Earth to oscillate in ways that we describe in Chapter 19. But fundamental changes in atmospheric or oceanic chemistry have not occurred during this time. Life has not had access to new energy-producing systems allowing new capabilities. Until such time as this happens (perhaps accelerated by human invention), we might expect that the Ediacaran–Cambrian revolution will continue to be played out in relatively minor innovations in body forms and functions.

Stimulating these minor innovations are the modest environmental changes hinted at in the preceding paragraph. These environmental changes, cyclic and stochastic, along with life's response to them, constitute the post-Cambrian history of Earth. In Chapter 19, we examine some of the cyclical changes. In the remainder of this chapter we focus on catastrophic change as a vehicle for "emptying ecosystems", allowing creatures with new innovations on the old body plans to gain ascendancy.

18.3 Mass extinction events in the Phanerozoic

Five mass extinction episodes occurred in the Phanerozoic eon. These events are identified in the fossil record by the apparently sudden disappearance of large numbers of families of organisms. The severity of extinction events is classified in terms of the fraction killed of families that leave reliable fossil remains. For example, shelly marine organisms are often a good marker of extinctions: a major late-Permian event killed off half of the shallow-water shelled families, and 80% of the species contained in total in those families. Figure 18.4 shows the number of families in certain marine and nonmarine phyla during the Phanerozoic; major extinction events are marked, but many smaller extinction events are reflected in dips in the number of families. The graphs show that such events can have a long-lasting effect on the diversity of living forms. The graph for marine organisms also shows that the initial production rate of different families at the start of the Cambrian, extrapolated to today, would have yielded 10^{27} families at present, instead of the roughly 2,000 actual living marine families. Clearly, both catastrophies and the saturation of ecological niches in between extinctions work to limit the variety of life.

The causes invoked for mass extinctions have been many and varied: drastic changes in climate such as major ice ages, large-scale volcanic eruptions, and impacts of large asteroids or comets are the most plausible candidates. The underlying causes of climate changes are themselves many and varied. We defer a discussion of cyclic or periodic climate-change processes to Chapter 19. Here we consider in detail the evidence for, and effects of, impacts. Many of the effects of volcanic eruptions are similar to those of impacts, and the focus on impacts is not intended to minimize the role of large-scale eruptions. However, only recently has the importance of impacts in terrestrial processes been recognized, a result of the reconnaissance of the solar system and its accompanying images of crater-pocked surfaces (Figures 7.3, 7.4). In particular, a compelling case has been made that the great extinction at the close of the Cretaceous, some 65 million years ago, was precipitated by a large impact.

18.4 Cretaceous–Tertiary extinction

In the time from the Cambrian to the Cretaceous period, a span of nearly half a billion years, animal phyla spread from the sea to continental environments, starting just shortly after the higher plants came onto the land. The vertebrate subphylum was represented first, primarily by amphibians and then by reptiles about 300 million years ago. Subsequent to the major Permian mass extinction, a subclass of reptiles called dinosaurs (Archosauria) diversified and occupied a large number of different niches on land and sea, equivalent roughly to those occupied today by mammals and birds. Mammals had developed by this time, about 250 million years ago in the Triassic, but were less successful than the dinosaurs and were restricted to species of small rodent-like creatures. Famous for the enormity of their size and diversity of forms and habits, dinosaurs were the dominant large animal on land and sea up to 65 million years ago.

At the end of the Cretaceous, a dramatic demarcation in the fossil record occurs, wherein 15% of the shallow-water marine families become extinct – including 80% of shallow-water invertebrate *species*. The dinosaurs disappear too, though the massive size of their fossilized skeletons make it difficult to conclude definitively that their disappearance was sudden. (The youngest dinosaur fossil yet found, a piece of a Triceratops, occurs just 13 cm below the upper end of the Cretaceous sediments – remarkably close). Above the Cretaceous sediments

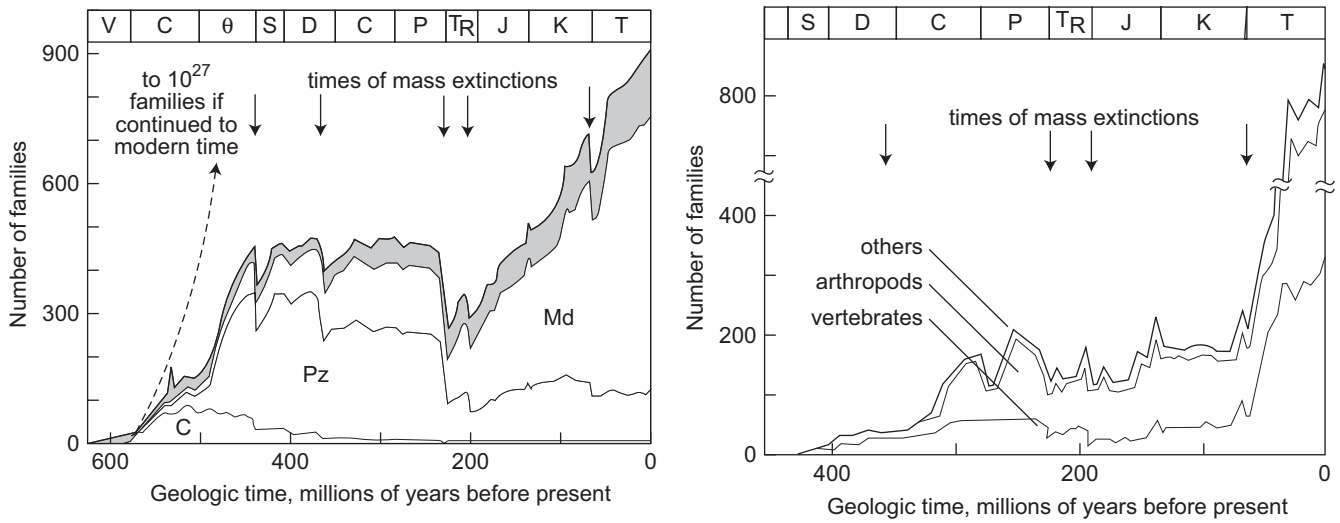


Figure 18.4 Number of biological families of (left) shelly marine fauna and (right) nonmarine animals over time. Dashed line extrapolates to later times the increase in number of families in the early Cambrian in the absence of extinctions. Arrows show times of mass extinctions. On the nonmarine graph, vertebrates and arthropods are broken out as well. Note that animals did not appear in nonmarine sediments, and hence were not on land, until almost 200 million years after the start of the Phanerozoic. The top portion of the graph labels the geologic periods using traditional symbols. From Milne *et al.* (1985).

lies the Tertiary, a time when mammals diversified and occupied the niches left empty by the demise of the dinosaurs.

The apparently contemporaneous disappearance of the dinosaurs, other land species, and large numbers of marine organisms as recorded in the rock record is called the *Cretaceous–Tertiary boundary extinction*, or K/T boundary event. "K" is the common geologists' symbol for the Cretaceous period of Earth history, based on the German word *Kreide* for Cretaceous. The K/T boundary in the geologic record worldwide is dated isotopically at 65 million years ago.

18.4.1 Boundary sediments

The dividing line at the K/T boundary is a thin (inches in extent) layer of clay. It has been identified in sediments worldwide. In addition to the sudden disappearance of many small marine organisms at the boundary, replaced in sediments above by more modern forms, the clay contains peculiar abundance anomalies. The platinum-group elements – iridium, osmium, gold, platinum, and others – more closely resemble abundances in meteorites than in the crust of the Earth. As we have discussed, the crust of the Earth is chemically differentiated and hence quite different in abundances relative to those of the bulk Earth. In particular, at the K/T boundary, iridium is more abundant than in normal crustal rocks. Other properties of the thin boundary clay layer or adjacent layers include the following items:

1. *Shocked quartz.* The impact of a large asteroidal or cometary fragment with the Earth's crust produces shocked grains of quartz with distinctive structure that allows their ready identification. These are seen in K/T boundary sediments around the world.
2. *Melt spherules.* Large impacts eject droplets of molten rock from the forming crater. These cool rapidly and solidify during their flight to form distinctive spherules. Their specific characteristics and occurrence over a large geographic area are diagnostic of an impact origin.
3. *Graphite.* Graphite, a form of carbon, and other evidence for burning is found at the K/T boundary in some parts of the world. Such burning would be expected from the large amount of debris lofted high into and even above the atmosphere: during re-entry this material would heat by friction (as does the Space Shuttle during its re-entry) and radiate this heat down to the surface. Distributed over a geographically large area by the impact, the molten debris would heat the air and the ground enough to ignite forests.
4. *Tidal wave action.* Impact into the ocean would generate large waves, with destructive effects upon reaching shore. There is evidence in some K/T boundary sediments adjacent to the Gulf of Mexico of patterns interpreted to be sudden wave action at the time the sediments were deposited.

18.4.2 Interpretation of the K/T boundary as an impact event

An asteroidal or cometary fragment roughly 10 km in size, striking Earth at high velocity can explain the characteristics of the K/T boundary layer and adjacent material. Such an impactor would gouge a crater roughly 200 km in diameter, blow a temporary hole of that size through the atmosphere, and fling dust into the upper atmosphere. The molten flying debris and pressure wave would burn and knock down trees across thousands of kilometers of land. Rock near the impact site would be "shock heated," changing its character to that seen, for example, in the shocked quartz of the K/T boundary. If the impact were into water, the resulting tidal wave would inundate adjacent land areas for hundreds of kilometers around. The material blown into the stratosphere (20 km or more above the surface) would circle the Earth; enough dust would be available

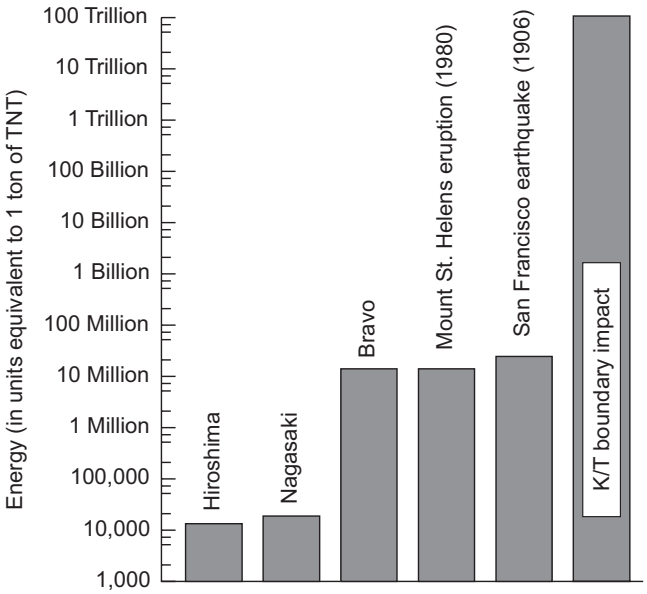


Figure 18.5 Comparison of the energy released in a putative K/T impact event with other energetic processes. Bravo is the largest US nuclear explosion. From Kring (1993).

to shroud the Earth in darkness for months. The dust, a mixture of impactor and crustal material, would fall onto continental and seafloor surfaces, carrying with it the chemical signature of the asteroid.

The general properties of the K/T boundary appear to be best explained by an impact. Volcanism involving magma from the deep interior would produce an iridium enhancement, but the ratios of iridium to the elements gold and osmium are different from those in meteorites and the K/T sediments. Fire fountaining in volcanoes can produce spherules, but usually on a local scale, and in basaltic eruptions; by contrast the K/T spherules are andesitic or dacitic. The arguments against volcanism as a cause of the K/T boundary phenomena do not rule out episodes of widespread volcanism as causing mass extinctions at other times in Earth's history; they pertain only to this particular event.

The energy released from such an impact is extraordinary, as shown in Figure 18.5. Over a million times more powerful than the Mount St. Helens eruption or the largest nuclear test explosion, such an impact has no rival with regard to anything experienced by humankind. However, serendipity allowed humans to view a similar impact into another planetary atmosphere, that of Jupiter, in 1994. In 1992, the orbit of Comet Shoemaker–Levy 9 was perturbed by a close pass to Jupiter and the comet itself was torn into two dozen fragments. Two years later these fragments collided with Jupiter's atmosphere. Although the largest pieces were a kilometer or less in size, the great gravity of Jupiter resulted in a much larger entry speed into the atmosphere of Jupiter than would have been typical at Earth. Since the energy of motion (kinetic energy) scales as the speed squared, this higher velocity was such as to make the energy of the biggest impacts similar to that of the K/T impactor. Clouds of dust the size of Earth, created by the comet fragments and by chemistry in the hot plume of rising gas after impact, were imaged by telescopes around the world and in space. Hubble Space Telescope in particular caught a spectacular glimpse of the mushroom clouds raised by the impact (Figure 18.6).

18.4.3 Biological effects of the impact

In addition to the direct effects of the impact, such as widespread forest fires and tidal waves, the plume of debris and smoke injected into Earth's upper troposphere and stratosphere would have a devastating effect on life through alteration of the climate. By physically blocking the rays of the Sun, the dust would cause the lower atmosphere and the surface of Earth to cool suddenly, and remain this way for weeks or months. Models suggest much of the continental area of Earth would have average daytime temperatures of only 10°C, much lower than the present average. In addition to the direct effects of cold on large animals and plants, the reduced sunlight would slow or shut down photosynthesis for up to a year, killing off large numbers of species dependent on various marine and continental food chains.

Analysis of the K/T boundary sediments suggests that the impact site might have been a seafloor carbonate platform, where the carbonate was deposited over long periods of time by shell-forming organisms. The amount of carbon dioxide released from the carbonate target on impact could have been very large, forcing the atmosphere to heat strongly after the cooling dust settled. Thus months of global winter might have been followed by years, decades, or more of global warming.

Another important effect of the impactor was on the chemistry of the atmosphere. As the impactor, or bolide, streaked through the atmosphere, it heated the air around it and caused chemical reactions to take place between the nitrogen and oxygen to produce nitrous oxides. Lightning does the same thing when it strikes, but the bolide could have produced enough that, when dissolved in water, the nitrous oxides and other synthesized compounds would have killed off large numbers of marine organisms. Calcium-shelled species could have had their shells dissolved by a change in the acidity of lake waters; such shelled organisms seemed to have suffered the most in terms of the fossil record of species extinctions. Sulfur oxides released from the seafloor would have converted into sulfuric acid in the stratosphere, creating acid rain that may have defoliated vegetation and altered water acidities over large areas.

Other effects of the impact may have been longer lasting. Tremendous amounts of water could have been injected into the stratosphere by the impact, reducing the ozone abundance, which then would have increased the incidence of cellular damage in surface organisms through elevated levels of uv radiation. Hydrogen sulfide produced in large quantities by worldwide decaying vegetation may have caused secondary food chain disruptions.

18.4.4 Chicxulub: the crater from the K/T impact event?

For an impact event, a crater of suitable size is the smoking gun. But for the K/T boundary, no such crater of suitable size was found in the decade after the impact hypothesis was first proposed. If the impact was in the ocean, the part of the crust containing the crater might have traveled to a subduction zone and

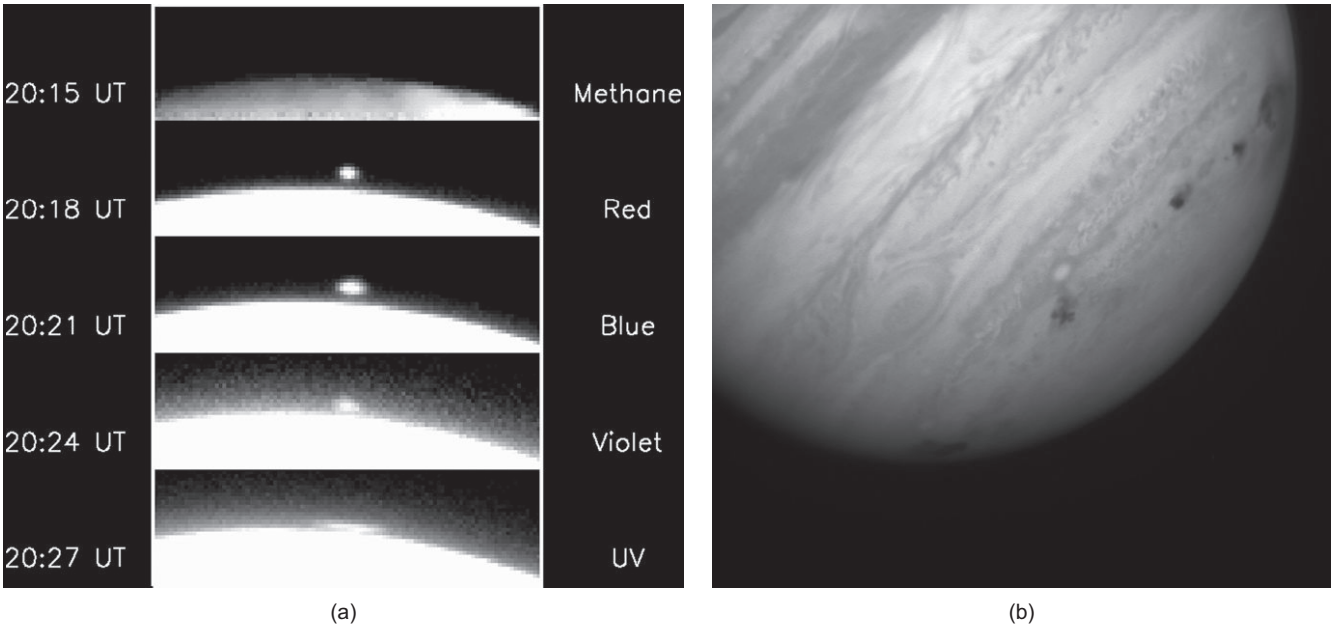


Figure 18.6 Hubble Space Telescope views of the 1994 impact of the Comet Shoemaker-Levy fragments into Jupiter. (a) Mushroom cloud rising over the Jovian limb. (b) Clouds of dark material from multiple impacts, each of which is larger in area than the Earth. Courtesy NASA/Space Telescope Science Institute.

been destroyed by subduction. However, careful detective work by a number of US, Canadian, and Mexican geologists finally identified a plausible candidate impact crater buried beneath part of the Yucatan Peninsula of Mexico and extending under surface sediments in the Gulf of Mexico (Figure 18.7). The buried crater, is called *Chicxulub* after the town that sits on the Yucatan Peninsula nearly over the center of the crater. Evidence supporting this site for the K/T impact event includes:

1. The thickness of deposited debris at the K/T boundary (*ejecta*) decreases with distance from Chicxulub, consistent with it being the source. Sites relatively close to the postulated crater show evidence for two layers: the low-energy debris thrown out and deposited locally and the high-energy material projected into a globe-circling layer.
2. The size of spherules and amount of shocked quartz is much larger in the Caribbean region than elsewhere. The chemistry of some of the spherules close to Chicxulub is high in calcium, consistent with impact into the calcium carbonate deposits associated with the Chicxulub site.
3. Studies of the geology of the Chicxulub site, including mapping of the density of the crust by measuring gravity variations, and mapping of iron concentration using magnetic measurements, are all consistent with a buried impact crater being located there.

Detailed studies of the site, many of which were conducted in the 1970s for oil exploration, reveal a circular structure roughly 180 kilometers in diameter, with several other rings interior and exterior to the main one. Shocked material in and around the site dates isotopically to 65 million years ago, although one drilling site in the crater yields an age a few hundred thousand years prior to the impact event. If this early age holds, then the crater is not the remnant of the impact that defined the K/T boundary

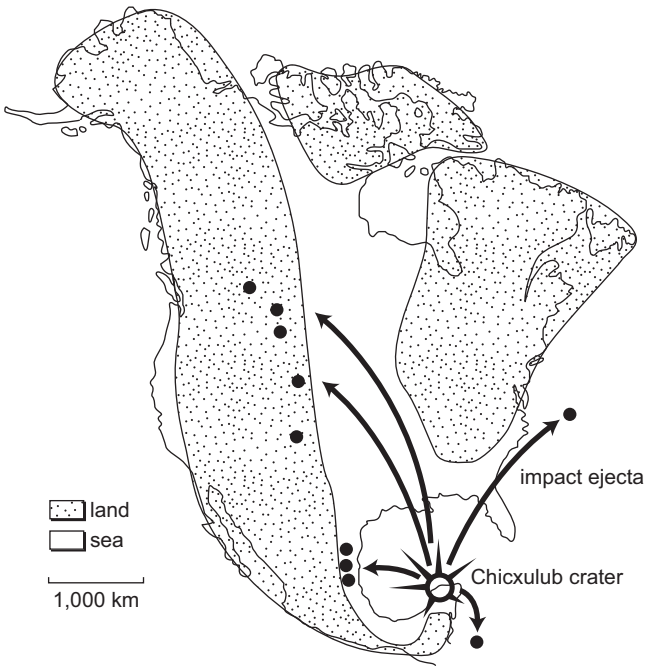


Figure 18.7 Map showing the location, off the Yucatan Peninsula of Mexico, of the Chicxulub crater. The present outline of the North American continent is shown. Shading indicates the approximate positions of North American continental areas that were above sea level during the late Cretaceous. Areas that are not shaded were under water, in part because of the very high sea levels of the late Cretaceous (see Chapter 19). Locations are labeled where impact ejecta associated with the crater have been collected. Courtesy of David Kring, Lunar and Planetary Institute, Houston.

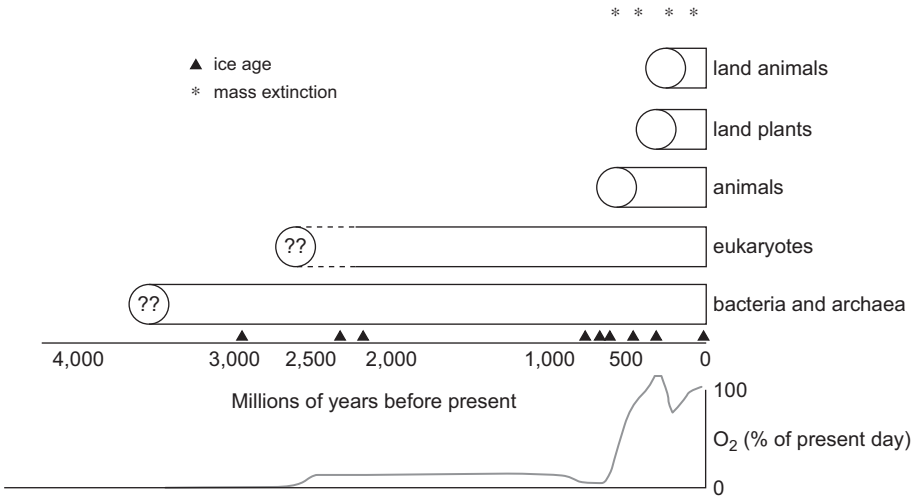


Figure 18.8 Appearance of forms of life in the fossil record, displayed against the occurrence of major glaciations (ice ages, triangles), and rise in oxygen (shown qualitatively in the lower panel). The major mass extinctions of the Phanerozoic are also shown.

through the iridium layer and extinctions. Several other craters of about the same age have also been discovered, so it is possible that the Earth was hit by a shower of debris over a time period that encompassed the K/T boundary. However, with the exception of one very large feature, whose impact origin is strongly disputed, these are all significantly smaller than Chicxulub. Thus, more work remains to assess in what way, if any, Chicxulub is connected to the K/T boundary event in the geologic record.

18.4.5 Impacts and other extinction events

Although several other major extinction events, which mark boundaries in the geologic time of the rock record, have been tentatively proposed as being associated with impacts, no records with the clarity of the K/T boundary exist for them. Hints of iridium enrichment, shocked material, or other evidence appear sporadically in other sedimentary layers, but never as abundantly as at the K/T boundary. Either the quality of the impact evidence at K/T is an anomaly, or impacts are not the primary cause of the bulk of Phanerozoic mass extinction events.

The characteristics of the Chicxulub impact site, with its carbon- and sulfur-laden marine sediments, probably ensured an unusually severe reaction on the part of Earth's atmosphere and biosphere to that impact event. It may well be that other impacts of similar magnitude at other times in the Phanerozoic did not have such a profound effect on life. The other major extinction events in the Phanerozoic may have had other causes, such as massive volcanism, or the more cyclic types of climate change to be discussed in Chapter 19.

Although this line of argument separates extinction events from the requirement that they be impact related, it does not explain why the sedimentary record does not show other impact events with the clarity of the K/T boundary event. Deep-ocean impacts, far from land, might well have lofted less dust worldwide – and the resulting ocean-floor crater might have been subducted and hence lost forever to discovery. Other impact events are preserved (usually poorly) on continental shields, but only a few rival Chicxulub in size. Evidently, the low frequency of large impacts in the Phanerozoic and the very active processes of erosion and subduction on Earth conspire to make the record of large impacts sparse indeed.

Many of the severe and sudden changes in climate and atmospheric chemistry induced by the bolide at the K/T boundary have their analogues in today's human activities, though on a less massive scale. Acid rain, enhancements to the greenhouse effect, and ozone depletion are with us today. A nuclear war could release enough dust through burning of cities to cool the surface of Earth and destroy agricultural food production for months or years. Direct human destruction of habitat, occurring now, may lead to loss of one-fourth of all species present on Earth over the coming decades. Such an extinction event, seen from the perspective of the fossil record, would be classified as intermediate or major along with the K/T event, the Permian event, and other great catastrophies of the Phanerozoic eon.

18.5 A global view of Earth's history so far

Figure 18.8 puts the Phanerozoic in perspective with the rest of the Earth's history. The Cambrian revolution in the appearance of animals is clearly seen, along with the "boring billion" of years between the assured appearance of eukaryotes and the Cambrian revolution. What is striking about this figure cannot be directly inferred from its content: one has to go back to Chapter 15 to recall that, within a billion to two billion years, Earth will lose its surface hydrosphere and hence complex life, if not all life, will become extinct. If Earth's situation is typical, and it takes billions of years after the formation of life for complex organisms to arise and become intelligent, then the probability of this happening on any given planet may be very small indeed. Is there time for another innovation akin to the Cambrian revolution to kick in before the brightening Sun closes the curtain on the history of life on Earth, or have we seen, in the record of fossils of trilobites and other wonderful creatures, all the innovation there really is in the genome?

Summary

The Phanerozoic eon began about 600 million years ago and is characterized by the diversification and global spread of multicellular organisms. While such organisms may have existed well before the start of the Phanerozoic, it was not until then different types of multicelled organisms rapidly diversified. The remarkable appearance of a variety of animal forms at the start of the Phanerozoic, the so-called Cambrian explosion, is a dramatic example of the process of evolution in action. Evolution is made possibly by the mutability of the genome in all organisms, but the nature of the changes that survive and propagate is shaped by natural selection: the effect of the environment on the organism. Without the two acting in tandem, the appearance of different and more complex forms might not have occurred. Evolution is not a slow, gradual process; species may remain stable for long periods of time, and only in the face of an event that isolates a breeding population might one see the appearance of a new species. For this reason and for the reason that the fossil record is an imperfect one, there are few cases of species change that are well documented in the fossil record, but those that are provide strong arguments in favor of evolution as the process by which new species appear. In the case of the Cambrian explosion, essentially all of the major animal branches or phyla appear at that time, along with some that did not survive to the present. Clues to the trigger for such a dramatic flowering of species may be found in the Ediacaran period that immediately preceded the Cambrian; a minor flowering of very primitive animal species took place at that time. What remains perplexing is the long delay between the development of eukaryotes and that of complex plants and animals. The delay may have to do with slow lengthening of the genome to allow for multicellularity, a sulfur-rich ocean, and one or more near-global glaciations that greatly restricted suitable habitats. Subsequent to the Cambrian revolution, much of the history of complex life has involved the interplay between ecosystem-emptying great extinctions, and the co-option of such ecosystems by new forms that diversified from classes of animals or plants that previously were unimportant. Thus the mammals were a relatively unimpressive class of animal until the dinosaurs, who occupied a much larger range of ecosystems, suffered extinction 65 million years ago. The cause of that great extinction remains controversial, but compelling evidence exists that a 10-km sized fragment of an asteroid struck the Earth, causing massive damage and climate change for a period of time. Unimportant in the overall history of the solar system as just another impact, the K/T boundary impactor paved the way for the diversification of mammals and hence, eventually, to ourselves.

Questions

1. Can you conceive of several alternative explanations for the lack of transitional forms in the fossil record? Explain why, logically, "absence of evidence" (of fossils) is not "evidence of absence" (of the evolutionary process).
2. Is the Ediacaran–Cambrian revolution an inevitable result of increasing genetic complexity? If so, what might you imagine could happen in a putative future revolution? Is such a revolution prohibited by external environmental conditions?
3. Using the formula for kinetic energy compare the amount of energy deposited by projectiles with radii of 1, 10, and 100 km, all moving at 10 km/sec. What happens to the energy if the speed is doubled? Assume the projectiles are spherical and have densities around that of rock (3 grams per cubic centimeter).
4. The concept that genome size must increase for more complex animals to arise seems to be contradicted by the observation that amphibians have a larger genome size than do all other types of animals. It is also contradicted by the fact that the human genome has half the number of genes that wheat does. Can you think of some other aspect of the genome that might determine the sophistication or complexity of an organism? (This may require a literature search.)

General reading

Gaidos, E. and Knoll, A. H. 2012. Our evolving planet: from the Dark Ages to an evolutionary renaissance. In *Frontiers of Astrobiology* (eds. C. Impey, J. Lunine and J. Funes). Cambridge University Press, Cambridge UK. In press.

Gale, J. 2009. *Astrobiology of Earth: The Emergence, Evolution and Future of Life on a Planet in Turmoil.* Oxford University Press, New York.

Margulis, L. and Sagan, D. 1986. *Microcosmos*. Summit Books, New York.

References

Cloud, P. 1988. *Oasis in Space: Earth History from the Beginning.* W. W. Norton, New York.

Eldredge, N. and Gould, S. J. 1972. Punctuated equilibria: an alternative to phyletic gradualism. In *Models in Paleontology* (T. J. M. Schopf, ed.). W. H. Freeman and Company, San Francisco, pp. 82–115.

Gaidos, E., Dubuc, T., Dunford, M. *et al.* 2007. The Precambrian emergence of animal life: a geobiological perspective. *Geobiology* DOI: 10.1111/j.1472-4669.2007.00125.x.

Gould, S. J. 1969. An evolutionary microcosm: Pleistocene and recent history of the land snail *P*. (*Poecilozonites*) in Bermuda. *Bulletin of the Museum of Comparative Zoology* **138**, 407–531.

Gould, S. J. 1985. *The Flamingo's Smile: Reflections in Natural History*. W. W. Norton, New York.

Keller, G., Adatte, T., Stinnesback, W. *et al.* 2004. Chicxulub impact predates the K-T boundary mass extinction. *Proceedings of the National Academy of Sciences of the USA* **101**, 3753–8.

Kring, D. A. 1993. The Chicxulub impact event and possible causes of K/T boundary extinctions. In *Proceedings of the First Annual Symposium of Fossils of Arizona* (D. Boaz and M. Dornan, eds). Mesa Southwest Museum and Southwest Paleontological Society, Mesa, Arizona, pp. 63–79.

Lyson, T. R., Bercovici, A., Chester, S. G. B., Sargis, E. J., Pearson, D., and Joyce, W. G. 2011. Dinosaur extinction: closing the 3 m gap. *Biology Letters* **7**, 925–8.

Milne, D., Raup, D., Billingham, J., Niklaus, K., and Padian, K. (eds) 1985. *The Evolution of Complex and Higher Organisms*. NASA SP-478. U.S. Government Printing Office, Washington, DC.

Vickery, A. C., Kring, D. A., and Melosh, H. J. 1992. Ejecta associated with large terrestrial impacts: implications for the Chicxulub impact and K/T boundary stratigraphy. *Lunar and Planetary Science* **XXIII**, 1473–4.

19

Climate change across the Phanerozoic

Introduction

The preceding chapter focused on singular events in the later history of the Earth – the flowering of multicellular complex organisms at the start of the Phanerozoic eon and the widespread extinction of species some 65 million years ago at the close of the Cretaceous period. Although these events stand out in their drama and the mystery of their causes, any understanding of the interactive history of life and Earth's environment cannot rest on their study. Throughout the Phanerozoic, and before, the relatively steady rhythms of plate tectonics brought continental masses together and then moved them apart, creating new seafloor and destroying old. The process of great landmasses moving around the planet must have had profound effects on the environment, and indeed this is seen to be the case in the geologic record.

This chapter begins by reconsidering plate tectonics with an eye to understanding the apparently cyclical creation and break up of multicontinent landmasses, or *supercontinents*. We consider the effects of such supercontinent cycles on the amount of volcanic activity, and hence atmospheric chemistry, on the ocean circulation patterns, on mountain building, and hence on the available area for storage of continental snow and ice deposits. Such considerations touch on a major theme of the latter portion of Earth history, the comings and goings of great ice ages. Finally, we draw our attention in detail to a particularly warm time in recent Earth history, the Cretaceous period. Ice free and showing much less drop in temperature from equator to pole than Earth experiences today, the Cretaceous has become a proving ground for climate modelers who seek to predict the amount and nature of global warming in humankind's future.

19.1 The supercontinent cycle

The ultimate causative agent of plate tectonics is the release of heat from Earth's interior through mantle convection, but the details of continental movement and seafloor subduction cannot be tied directly to the interior convective patterns, at least based on computer models simulating those deep motions. Instead, the surface patterns of plate motion depend upon several things visualized in Figure 19.1: the age and density of the oceanic crust, collisions between continents, and the deflection of mantle heat sources by piled-up supercontinental masses.

Oceanic crust newly created at mid-ocean ridges is hot, and hence relatively buoyant. As this crust is displaced by yet younger crust, it rolls laterally away from the ridge, cooling as it does. Cooler crust contracts, and becomes denser. If the older oceanic crust does not encounter a pre-existing subduction zone, forcing it under, it eventually will cool and densify enough to sink spontaneously, creating a new subduction zone. Evidence from magnetic reversals on the seafloor (Chapter 9) that no portion of oceanic crust is older than 200 million years is buttressed by computer models suggesting that beyond that age the ocean crust is indeed too dense to be supported by the asthenospheric part of the mantle. (We exclude oceanic crust thrust up onto continents as "ophiolites".)

Continental collisions are self-explanatory: because continental crust is buoyant at any age, collisions between continental landmasses on adjacent plates force the directions of plate motions to shift. Strong compression during such collisions raises mountain ranges, such as the Tibetan Plateau (with Mt. Everest), raised by the current collision of India with Asia. Furthermore, as bigger aggregations of continents build, heat flow from the mantle is inhibited by the thick crusts and insulating properties of these buoyant masses. As a result, heat flow elsewhere may increase, precipitating new oceanic ridges, or may eventually rift the continents apart again. The idea that plate motions on long timescales have a cyclical character defined

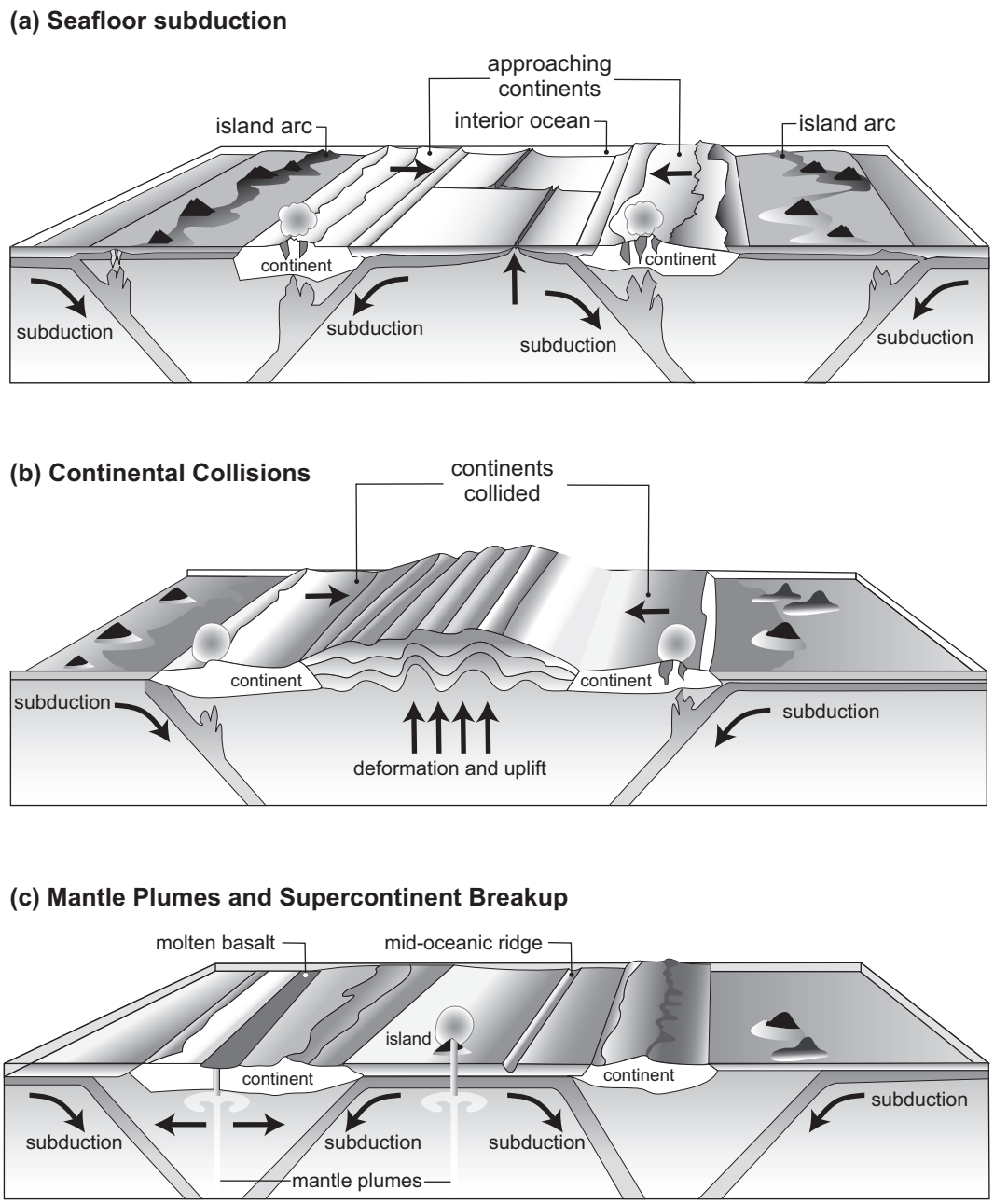


Figure 19.1 Three processes important in the determination of plate motions: (a) subduction of cold, dense ocean crust; (b) collisions between buoyant continental masses; (c) effect of thick continental crust on heat flow from the mantle. In panel (c), a mantle plume has developed beneath the supercontinent on the left, encouraging break up.

by continental collisions is suggested by the tracking of plate motions as far into the past as feasible, perhaps a billion years or more. Originally proposed by Toronto geophysicist J. Tuzo Wilson and refined by others, the supercontinent cycle goes as follows:

1. The continents are collected together in a single amalgamated mass (a supercontinent), surrounded by a global ocean (a universal ocean).
2. Mantle heat is trapped beneath the supercontinental crust; temperatures rise within the crust, causing expansion, arching, and fracturing of the supercontinent. Additionally, the spin of Earth puts a small additional stress on the supercontinent, which sits like a raised pimple above the ocean floor and hence is subject to a higher centrifugal force than the seafloor crust.
3. Rifting of the supercontinental mass occurs along one or several lines. Mantle material rising up in the space between the newly fragmented continents partially melts, forming oceanic crust along a new mid-ocean ridge in a growing "inland" ocean. As new seafloor is created, the continents spread apart, the boundary between continent and seafloor being a tectonically quiescent *passive margin*. In the universal ocean surrounding the exterior margins of the continents,

subduction zones at continental margins and elsewhere consume seafloor, shrinking the universal ocean.

4. Seafloor at the continental margin of the new ocean becomes older and colder until finally buoyancy is lost and subduction begins. Subduction halts or redirects the growth of the new ocean. Continental masses no longer spread outward but may begin to converge again until collisions recreate a single supercontinent.

Figure 9.10, showing the motion of the continents over the past 200 million years, illustrates the first half of the supercontinent cycle. The break up of the last supercontinent, Pangaea, initiated the opening of the Atlantic Ocean and the shrinking of the Pacific. The margins of the Atlantic do not contain subduction zones, but instead are passive boundaries with the surrounding continents. In contrast, the continental margins in the Pacific are sites of active subduction zones or, where lateral motion is taking place, transform faults. The earthquakes and volcanic eruptions along the Pacific ring of fire stand in stark contrast to the quiet of the Atlantic region. The supercontinent breaks up not once, but several times, until the current number of separate landmasses is reached. Eventually, perhaps in a few tens of millions of years or less, the Atlantic will develop subduction zones as cooling ocean crust loses buoyancy. Tectonic activity will develop along the Eastern seaboard of the United States, Western Europe, and West Africa. The expansion of the Atlantic will end and the continents will eventually collide back together to form a single landmass.

Reconstructions of early plate tectonic cycles support the notion that previous supercontinents existed, the one prior to Pangaea rifting apart perhaps 700 million to 800 million years ago. Tenuous evidence for an earlier episode also exists in rocks a half-billion years older still. So, supercontinents break up and then come back together every half-billion years or so, perhaps as far back as the end of the Archean when enough continental mass existed to influence the motion of the crustal plates.

19.2 Effects of continental break ups and collisions

The separation and collision of continents does more than just alter the geographic map of the world over time; changes in continental positions and possible accompanying pulses of geologic activity play roles in altering climate. These effects continue to be an active area of research, and a detailed correspondence between plate positions and possible ancillary events in the geologic record remains elusive. However, several potential effects can be identified.

Mountain building is associated both with the expansion of continental masses away from the supercontinents and with subsequent collisions. Interior mountain chains and highland plateaus result from continents colliding with each other; mountain chains along the exterior of a continent are built up by volcanism associated with the subduction of ocean floor beneath the edge of the continent. In either case, the build up of new continental highlands produces more land area for ice accumulation, with effects that we discuss in section 19.5.

Volcanism associated with the formation of new subduction zones along continental margins as well as in the seafloor exterior to the diverging continents puts large amounts of ash, aerosols, and greenhouse gases into the atmosphere. Like large impacts, the initial effect is a cooling as atmospheric aerosols reflect or absorb some sunlight. Eventually, the aerosols drop out, but the carbon dioxide and other greenhouse gases added to the atmosphere remain for much longer and contribute to a hotter climate. Volcanic gases and ash added to lakes and seas change the acidity of the waters, altering their suitability for adapted organisms.

Volcanic episodes are not restricted to the continental margins; a surge of eruptive activity associated with the initial rifting of a supercontinent may have dramatic climate effects as well. A massive extrusion of lava over a 517,000 square kilometer (200,000-square-mile) region, the so-called Deccan Trap lava flood in India, occurred some 65 million years ago, associated with rifting away of part of the continent. Ancillary effects of the eruptions might have played a role in climate change and possibly extinctions near the K/T boundary, additional to (or, some argue, in place of) a large asteroid impact.

Changing continental positions have two primary effects on climate. First, the drift of continental fragments toward higher latitudes than those occupied by the supercontinents, which seem to have had their geographic centers at low latitudes, allows more snow and ice accumulation to take place. High-latitude continents are better accumulators of snow and ice than are high-latitude seas, primarily because continental areas have elevated terrains. Second, as continents drift, ocean currents, which transport warm and cold ocean water over vast distances, shift in their strength and direction. The so-called North Atlantic deep water, an area of sinking salty water that strongly moderates Europe's climate, is shaped in large measure by the North American continental margin. (The role of this major ocean feature in climate is discussed in Chapters 21 and 22.)

If indeed the motion of tectonic plates plays a role in determining Earth's climate, such modulation should be present in the geologic record. And, so it is, in the form of epochs of ice ages that have occurred a number of times over the history of the Earth.

19.3 Evidence of ice ages on Earth

Ice ages is a colloquial term for glacial episodes – times in Earth's history when glaciers covered large areas of the continents, down to mid-latitude regions and hence much farther equatorward than today. "Snowball Earths" refer to extreme episodes wherein ice may have extended most of the way to the equator. Glaciers, year-round sheets of ice and entrained rocks of all sizes from grains to huge boulders, leave characteristic signatures as they advance across the landscape and then break up. (Few glaciers recede large distances intact.) These features are distinct from the erosive effects of liquid water because of the very different mechanical properties of liquid water and ice.

Glaciers carve out U-shaped valleys and bowl-shaped *cirque* basins in mountainous terrain. On a continent-wide scale, the advance of glaciers with their embedded rocks scratch and striate the surface. Debris pushed ahead of glaciers and abandoned

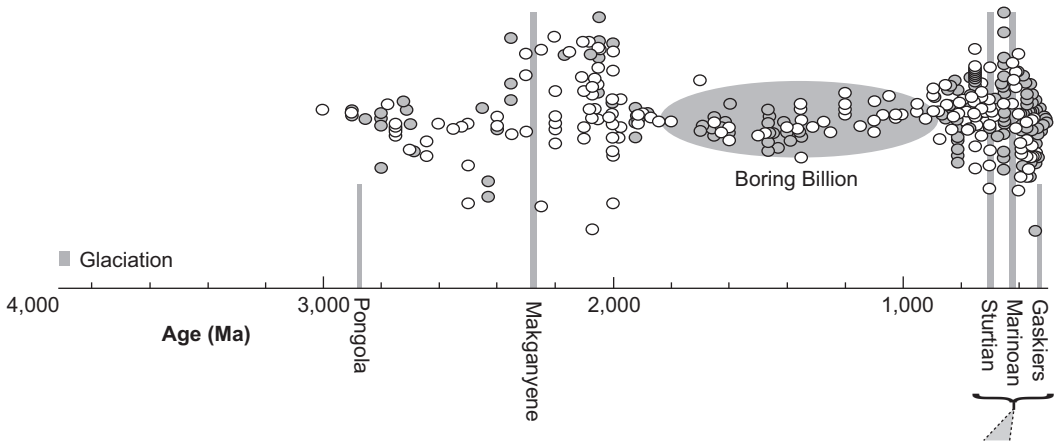


Figure 19.2 Timeline of ice ages. Bars correspond to times in Earth's history during which widespread glaciation occurred, based on geologic data from a number of locations around the globe. Times are marked in millions of years before present: thus, "1,000 Ma" is a billion years ago. See Chapter 18 for a discussion of the "boring billion."

when the glaciers vanish creates the undulating *moraine* terrains. The sheer weight of ice sheets that rise 3,000 meters above the surrounding terrain depresses the upper continental crust; as the glaciers disappear and the land rebounds, lines of stress called *strandlines* appear over large areas. On a small scale, glaciers do not sort and round rocks the way streams do – poorly sorted angular material is more characteristic of glacial debris. In some rare cases, freezing muds may capture the imprints of ice crystals at the base of the glaciers.

Such geologic signatures (and others) of glacial activity are present at sites where glaciers still exist – or did in historical times – and amply over the broad northern continental ranges affected by the glaciations of the past million years. To adduce the existence of much earlier ice ages, back billions of years, is a much more difficult proposition. Perhaps the extreme case of this is the attempt to infer glacial epochs on Mars, as described in Chapter 15, where geologic processes have been dominated by impacts and some volcanism, with ancient episodes of water erosion. Since only a tiny part of Mars has been examined by landed instruments, the search for glacial features is limited to orbital surveys, and hence only large-scale features serve for now as the (rather controversial) evidence for sheets of ice sometime in Mars' past.

On Earth, at least, the rocks can be examined at close range. Ancient rock strata preserved in the older shields of the continents must be examined for the small-scale evidence of glacial action; large-scale glacial terrains from ancient ice ages have been largely erased by subsequent tectonic and erosive processes. The most common and diagnostic indicators of the existence of ancient glaciers are rock surfaces that are polished and striated, pebbles with a characteristic shape associated with glacial scouring, and agglomerations of large angular rock fragments in a fine-grained matrix.

Other signatures in the sedimentary rock record have been used to infer several major episodes of glaciation over Earth's history. Oceanic reversion during snowball Earth episodes to anoxic conditions creates layers of unusually young banded iron formations, dated to 750 million years ago (Chapter 17). A steep drop in ^{13}C to ^{12}C in carbonate layers suggest depressed biological productivity and the onset of cold times (Figure 19.2).

19.4 Causes of the ice ages

19.4.1 Positive feedbacks in the basic climate system

Widespread continental glaciation represents a distinct state of the complicated physical system comprising Earth's atmosphere, oceans, and continents. As with many complicated, nonlinear physical systems, a series of small changes may push the system into an entirely different state, as positive feedbacks amplify the small perturbations. Continental ice cover is a good example. Adding ice sheets to a continent, for whatever reason, raises the *albedo* or reflectivity of the surface, ensuring that less sunlight is absorbed by the ground, and hence less energy is reradiated as infrared photons back into the atmosphere. The contribution to the annual mean temperature and the atmospheric heat budget of Earth is less from regions that become ice covered, and global temperatures drop. This encourages more ice to form at even lower latitudes (on both land and oceans) and the system is driven toward a state in which large areas of Earth are covered in ice.

The triggers for such ice ages remain somewhat controversial. Clearly one trigger is the movement of continents, split off from a single supercontinent, toward higher latitudes. This drift puts more landmass in regions where cold climate allows ice accumulation. The production of mountain ranges associated with high-latitude continental collisions, collisions of continents with island arcs, or subduction zones pushes continental landmass to higher altitudes, encouraging further ice accumulation. The evidence for Proterozoic and Phanerozoic plate tectonic cycles of continental assemblage into supercontinents, followed by break up, is strong. Although correspondence between past ice ages and dispersal of continents cannot be made confidently because of uncertainties in the ages of both and in the timing of the onset of glaciations relative to continental positions, it is a plausible connection.

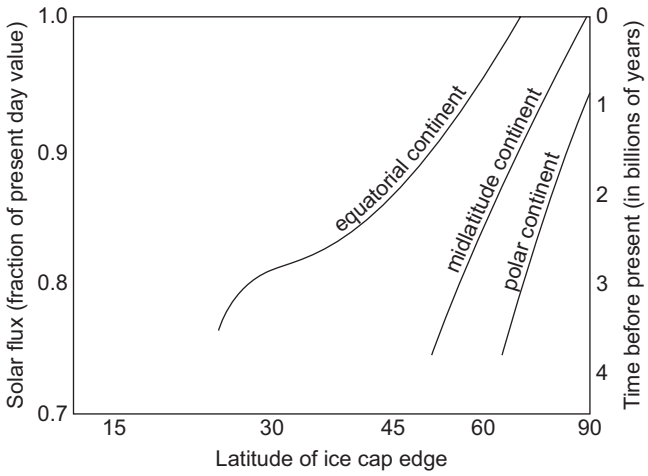


Figure 19.3 Example of the possible effect of continental positions on the severity of past ice ages. A model was constructed by University of Michigan scientists H. Marshall, J. C. G. Walker, and W. R. Kuhn of the balance between carbon dioxide consumption by weathering and release by volcanism and metamorphic heating. The weathering rate was varied depending on the latitude of an assumed single supercontinent. An equatorial supercontinent, receiving essentially all of its precipitation as rain, allows carbon dioxide to be consumed more quickly than does a near-polar supercontinent that receives its precipitation in the form of snowfall. The graph shows the lowest-latitude limit of ice sheets for three different model supercontinents at various times in Earth's history, corresponding to different values of the solar luminosity. At no time does the ice reach completely to the equator, but ice ages, once begun, are less severe when continents are confined to high latitudes. Adapted from Marshall *et al.* (1988).

19.4.2 Negative feedbacks in the climate system

In practice, negative feedbacks prevent Earth from going to a completely, permanently, ice-covered state. During ice ages, less precipitation occurs in the form of rainfall, and hence less erosion and removal of atmospheric carbon dioxide to the seafloor (as carbonates) occurs. This effect is accentuated if continental masses are at high latitudes, where precipitation is almost all snow and hence erosion is less effective (Figure 19.3). The volcanic outgassing of carbon dioxide previously subducted as carbonates continues regardless of the carbonate production rate so that, during the ice age, there is a net tendency of carbon dioxide to increase. This in turn increases the infrared opacity of the atmosphere, enhancing the greenhouse warming and eventually offsetting or ending an ice age.

The negative feedback associated with the resupply of carbon dioxide and other volatiles to the atmosphere distinguishes Earth from Mars. Mars has been in a perpetual ice age since early in its history, punctuated perhaps by only the briefest of episodes of running water. As described in Chapter 15, the absence of plate tectonics, the relative ease with which the atmosphere could escape to space, and the more distant Sun all played important roles in shunting the Martian climate to this state. Important here is the recognition that, although Earth's climate is not constant, but instead oscillates between warm and cold extremes, the feedbacks afforded by tectonic and other processes have kept these oscillations small enough that the basic state of stable liquid water is retained.

19.4.3 Additional influences on global glaciation

Other effects act on the extent and duration of glaciation but the direction and magnitude of each are harder to quantify. The positions of the continents determine in part the pattern of ocean currents that transport warm equatorial seawater to higher latitudes. The presence of high-latitude continents and high-altitude ice sheets alters the patterns of storm systems, hence affecting timing and amounts of rainfall and snowfall. Build up of mountain ranges and high plateaus also might increase the rate of weathering and subsequent loss of atmospheric carbon dioxide. Causes external to Earth may trigger ice ages as well. Early in Earth's history, the Sun's lower luminosity would have made it easier for Earth to slip into ice ages. In fact, absent the enhanced carbon dioxide abundance postulated for the Archean and Proterozoic atmospheres (Chapter 14), Earth would have been in a continuous ice age that could have thwarted the establishment and development of widespread life. Temporary dips in the Sun's luminosity later in Earth's history, or passage of the solar system through dusty molecular clouds, attenuating the sunlight reaching Earth, cannot be ruled out either as sources of cold episodes.

19.5 Cretaceous climate

The mid-Cretaceous, from roughly 100 million years ago to its conclusion 65 million years ago, appears to have been characterized by an Earth with no permanent ice caps, equatorial mean annual temperatures slightly higher than today, and polar-cap mean annual temperatures 40° to 60°C higher than today. Such a world would look from space much different than our present Earth, with the Arctic and Antarctic ice caps not present. It also would have been a far different place to live, with little variation in climate from the equator to high latitudes. It represents an extreme in climate, opposite to that of the deep global glaciations, and which can be studied in detail because it occurred recently in Earth history. Understanding this last warm time in Earth history is therefore a priority among climatologists, who also see in the Cretaceous a guide to the possible effects of human-induced global warming.

19.5.1 Evidence for the Cretaceous climate pattern

The following constraints exist on the Cretaceous climate:

1. *Isotopic data*. Stable isotope ratios, primarily ^{18}O to ^{16}O, are available for a number of sediments from Cretaceous times that were formed in equatorial and mid-latitude seas. By choosing sediments characteristic of both deep-sea and shallow-sea environments, it is possible to get a profile of ocean temperatures with depth, as well as latitude (Chapter 6).
2. *Fossil organisms*. Plate tectonic motions have carried continents far in the 100 million years since the mid-Cretaceous. It is possible to reconstruct the pattern of continents, which

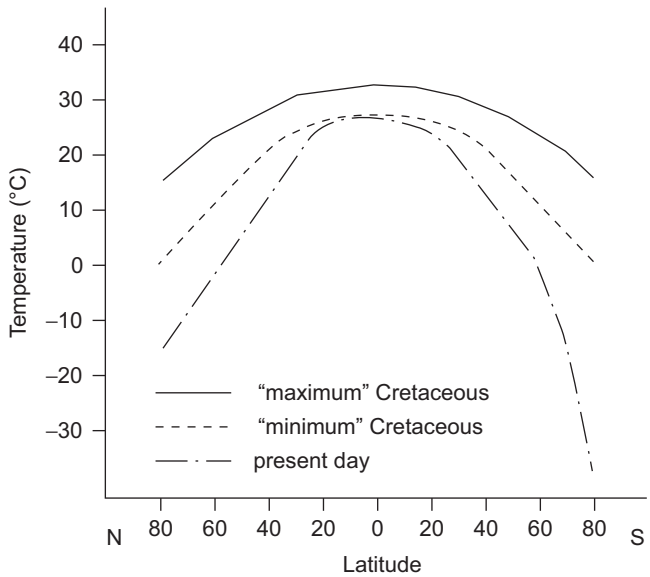


Figure 19.4 Estimated limits on temperature in the Cretaceous as a function of latitude. The plausible range (maximum and minimum) of annually averaged temperature at each latitude is shown, along with the value for the present-day Earth. From Barron *et al.* (1995).

then permits the location of Cretaceous fossils according to latitude to be determined. A number of fossils indicate equable climate to the poles at that time. Coral reef and carbonate formations extended 5° to 15° of latitude poleward of their current limits, because of warmer conditions. Fossil alligator and crocodile remains indicate that these tropical creatures lived at latitudes up to 60° north and south in Cretaceous times. Other fauna support this pattern; fossils of cold-water species are absent from the Cretaceous sedimentary record, and diverse numbers of warm-water species are present at high latitudes.

3. *Geology*. Glaciers are a primary force for erosion at high latitudes and high altitudes on Earth today. Yet the key patterns revealing glacial erosion are missing from Cretaceous rock formations that were at high latitudes. Some temporary ice may have formed during parts of the year during the Cretaceous, but year-round ice is largely ruled out by such findings.

The constraints on temperature provided by the range of evidence presented here are summarized in Figure 19.4 as annual mean temperature as a function of latitude during the Cretaceous. Two estimates based on the data – lower and upper limits – are compared with the present annual mean temperature at each latitude. There are several interesting effects: the global annual mean temperature in the Cretaceous was 6° to 14°C higher than today. The annual mean temperature at the equator was 0° to 5°C higher; the polar mean temperature was as much as 60°C higher. Instead of the 41°C equator-to-pole contrast that we see today, the contrast in the Cretaceous was only 17° to 26°C. Permanent ice and widespread seasonal ice were absent from Earth at that time.

Such warmth exceeds by a large amount the visions of the human-induced global warming predicted by computer models discussed in Chapter 22. To be able to reproduce such a different climate with computer models developed to predict weather today is clearly of keen scientific interest because such an exercise stretches the physical regimes under which such models have been fine-tuned.

19.5.2 Plate tectonic effects on Cretaceous climate change

Although the break up of the supercontinent Pangaea began in the Jurassic, the Cretaceous Earth still had most of its continental landmass at low and mid-latitudes. With little land available near the poles, ice accumulation was difficult. The overall reflectivity of Earth was therefore lower than at present, allowing more sunlight to be absorbed and encouraging warmer conditions.

But tectonic effects on the Cretaceous climate were more complex than simple land distribution implies. As the supercontinental bottleneck was broken, plate spreading rates were probably fairly high. Very active seafloor spreading brought relatively hot, puffed-up crust rapidly away from mid-ocean ridges. This, along with the absence of ice on the continents, implied a very high sea level, and water flowed over the continental lowlands to form vast inland seas. In consequence, the area of exposed land in the Cretaceous may have been only 60 to 70% that at present. These inland seas absorbed more sunlight than did the dry land, and may have been more important than the absence of ice in heating Earth's surface. Further, the inland seas were, on average, warmer than the ocean and probably helped to maintain mild sea-surface conditions at high latitudes through exchange of water with the ocean.

The spreading apart of Pangaea was a time of less mountain building, because continental collisions were minimal. Less mountain building meant less land area at high altitudes. The lower mean altitude may have implied less snow on the midlatitude continents, buttressing the effect of having little landmass near the polar regions. With fewer massive mountain ranges, as well as a higher sea level, the amount of continental surface area available for weathering may have been less than at present, leading to a lower rate of removal of carbon dioxide from the atmosphere. Also, faster plate tectonic recycling of the crust could have accelerated the rate of production of carbon dioxide from subducted carbonates, and injection of the gas into the atmosphere through greater volcanic activity.

19.5.3 Additional important effects on Cretaceous climate

Ocean currents. The broad universal ocean undoubtedly had a different pattern of ocean currents than today. Less continental land area was affected by such currents than today because a single landmass has less coastline than the same mass fragmented, which could have led to more severe latitudinal variations in continental weather. As the continents broke up in the Cretaceous, currents of water in the new Atlantic Ocean changed this pattern substantially.

Water vapor and clouds. Increased temperature of the oceans increases abundance of water vapor in the atmosphere, which increases the greenhouse heating. It might also increase the cloud cover of Earth, which can add to or subtract from the

heating, depending upon the thickness and geographic distribution of clouds. The effect of increased temperature on cloudiness, however, is very uncertain – for example, in the tropics, increased heating might lead to a greater preponderance of convective clouds (cumulus and thunderstorms), which create areas of locally heavy cloud but leave some of the sky cloud free.

19.5.4 Causes for climate change that probably are not important in the Cretaceous

There are other possible causes of climate change that cannot be directly ruled out but are either less likely to be relevant, or somewhat arbitrary in the way they must be invoked.

Solar output. Because the Sun has been heating up over time, we do not expect this trend to explain the relative difference between the Cretaceous and the present climate; the effect works the wrong way. Astrophysicists have suggested ways the Sun might brighten temporarily, and such a brightening could have triggered a warming, but it is impossible to determine whether the brightening timescale is commensurate with the duration of the warm period (some tens of millions of years).

Orbital variation. The variations in the orbit and tilt of Earth, described in section 19.8, occur on timescales much shorter than those required to explain the Cretaceous warmth.

Galactic effects. Passage of Earth through dusty clouds in the galaxy cools Earth rather than warms it. Perhaps we are in such a cloud now, and were not 100 million years ago. However, the magnitude of the cooling from the Cretaceous to present, and its gradual long-term nature, are hard to explain given what we know of the properties of such clouds.

19.5.5 Model for the warm Cretaceous

Scientists have used computer models to predict changes in the present Earth's climate on timescales of days, weeks, months, years, and decades. Such a computer model was adapted by E. Barron (Pennsylvania State University) and colleagues at the National Center for Atmospheric Research in Boulder to simulate the Cretaceous climate. The model simulates the atmospheric greenhouse effect discussed in Chapter 14, along with the transportation of heat in the oceans. We discuss this and models like it in much more detail in Chapter 22, where we consider concerns about present-day global warming.

The first test for the model was to change the positions of the continents to correspond to Cretaceous times without changing anything else. This produced only a fraction of the temperature increase over the present-day climate required to explain the Cretaceous warm period. Adding four times the present carbon dioxide abundance to the atmosphere enhanced the atmospheric temperature at the poles to close to the values inferred from the data, but then the equatorial temperatures were too high.

It appears that to explain the temperature pattern shown in Figure 19.4, there must have been enhanced transport of heat from the equator to the poles in the Cretaceous compared to the present. It is hard to make the atmosphere in the model transport the heat required, because a smaller temperature contrast from equator to pole actually means less efficient heat transport: The oceans must do the job (Figure 19.5). It is possible that the ocean circulation in the Cretaceous was organized in such a way as to promote very efficient transport of heat from equator to pole; computer models only recently have gained the sophistication to explore this possibility in detail.

It appears from the computations done to date that the most important differences between today's world and the Cretaceous that determined the warmer climate are (i) the pattern of continents, which was more consolidated toward equatorial latitudes in the Cretaceous; (ii) enhanced Cretaceous ocean circulation from equator to pole; (iii) enhanced Cretaceous carbon dioxide levels. In today's world, human activities have an effect only on (iii). Until more accurate representations of the roles of clouds, precipitation, and other effects can be included in the models (Chapter 22), these conclusions must be regarded as tentative.

19.6 The great Tertiary cool down

The impact event that destroyed much of the Cretaceous fauna apparently did not have a long-term effect on climate, because the early Tertiary was similar to the late Cretaceous with respect to climate. Indeed, the Eocene may have been as warm or warmer than the Cretaceous. As Figure 19.6 shows, however, by the mid-Tertiary the climate was cooling down, with some oscillations; by 2 million years ago global temperatures were cooler than at any time in the previous half billion years. Isotopic data on climate are excellent for this time, as the level of detail in the figure indicates.

Excursions in the Tertiary climate may be associated with increased volcanism, rapid changes in plate tectonic patterns (for example, India collides with the Asian continent in the Eocene), large variations in the Sun's luminosity, or even additional impact events such as the Cretaceous–Tertiary event (but smaller). Lesser extinction events occur in the late Eocene and in the Pliocene; one or both could be associated with climate shifts triggered by volcanic, tectonic, or impact events. The causes behind various swings and the overall cooling in the Tertiary part of the climate record remain poorly understood. Perhaps the climate record simply reflects the progressive departure of the tectonic and atmospheric states away from the Cretaceous condition of ice-free continents and high carbon dioxide content. The increasing ice coverage of the high-latitude continental areas, enabled by plate motion, reinforced the slow cooling, punctuated by occasional warmings of uncertain origin.

It is unsatisfactory not to have a specific mechanism for the cooling and decline of carbon dioxide, and a dramatic one has been offered from the observation that, roughly 40 million years ago, the crustal plate carrying the Indian subcontinent collided with the massive Asian continent. Since that time, India has continued to plow into Asia to build up the enormous Tibetan Plateau, location of the world's highest mountains. M. Raymo of MIT and colleagues W. Ruddiman (U. Virginia) and P. Froelich (Georgia Institute of Technology) have proposed that the continued build up of the Tibetan Plateau to the present has increased weathering and loss of carbon dioxide from the atmosphere. The presence of the plateau forces moist winds from the Indian ocean to rise and produce prodigious amounts of rain, which enhance weathering rates as well as feed eight of the Earth's large rivers, which in turn carry hydrogen carbonates and other weathering products to the sea. Additionally, the presence of a

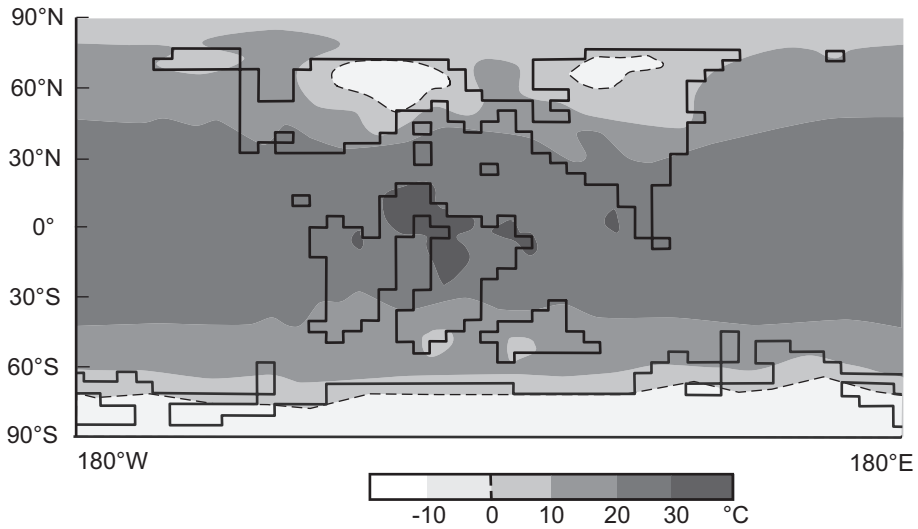


Figure 19.5 Predictions for annually averaged temperatures in the Cretaceous from a computer climate model. The results are displayed on a map of the world with the rough outlines of the continents as they would have appeared in the Cretaceous. Shading shows the annually averaged surface temperature, with a key at the bottom, in Celsius. The model has four times the present atmospheric carbon dioxide value and four times the present-day oceanic transport of heat from equator to poles; it satisfies the temperature constraints shown in Figure 19.4. From Barron *et al.* (1995).

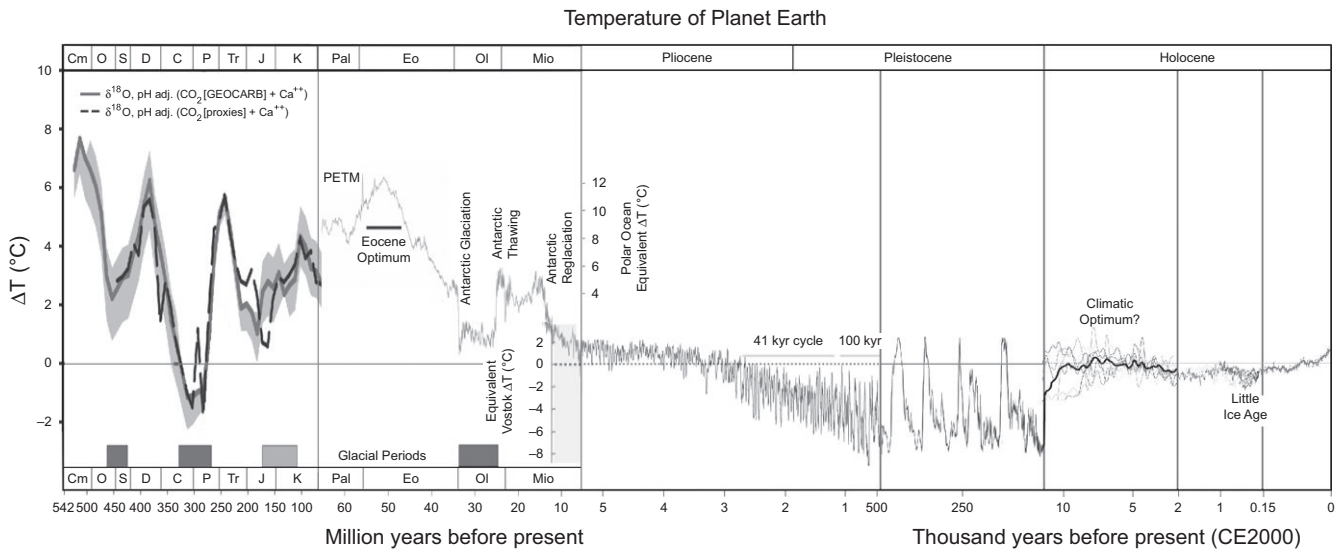


Figure 19.6 Summary temperature and precipitation for the past half-billion years of Earth history, based on isotopic and other indicators. Times before present and geologic periods are given on the bottom and top of the graph. General times of very low temperatures, and possible ice ages, are labeled at the bottom. Note the scale progressively stretches out toward recent times, reflecting the better resolution of more recent data. Vostok temperature scale refers to data derived from an Antarctic ice core, and the global temperature change is assumed to have been half the polar temperature change. Figure created by Robert Rohde from various data sources (Rhode, 2011).

large plateau with steep slopes increases the surface area of rock available for weathering, relative to a low flat plain. Computer simulations suggest that these results of the rise of the plateau have significantly increased the rate of removal of carbon dioxide from the atmosphere in the post-Cretaceous world. Indeed, the plateau may be so effective that only the negative feedback of an increasingly colder climate has prevented an essentially total and catastrophic removal of the carbon dioxide.

Figure 19.6 shows that the progressive decrease in temperature toward the present shifts suddenly to very dramatic oscillations in temperature beginning early in the penultimate geologic epoch, the Pleistocene. These climate oscillations are characteristic of the ice age that continues to the present. Readers may be surprised that our time is identified as such; however, the ice age epoch in which we live is characterized by long stretches of glacial conditions punctuated by shorter intervals, only one-tenth as long as the glacials, of warm interglacial climate such as the current Holocene. Prior to the onset of glacial oscillations, the slowly cooling climate led to conditions in which mammals flourished, having filled most of the ecological niches vacated by the dinosaurs. Only the air remained the domain of dinosaurs or, more precisely, their close descendants, the birds. The Eocene,

Oligocene, and Miocene epochs, stretching from 56 million to 5 million years ago, are really the golden age of mammals, with many more wonderful species of large mammals than humans have ever been privileged to see.

19.7 Causes of the Pleistocene ice age and its oscillations

The onset of the Pleistocene ice age was the end result of progressive cooling over the 50 million years prior to it. With enough continental landmass at high latitudes, low carbon dioxide in the atmosphere, and other properties that may not show in the geological record (for example, enhanced ocean mixing leading to more marine planktonic mass and higher consequent uptake of carbon dioxide), the climate system shifted abruptly to a state of widespread continental glaciation. There is more to the Pleistocene story than the glaciation itself, and that is the oscillatory nature of the climate. Long spans of continental glaciation, ranging from 40,000 years early in the Pleistocene to 100,000 years in the later cycles, are punctuated by interglacials of roughly 10,000 to 20,000 years characterized by warmer (or in some cases, unstable and rapidly shifting) climate (Figure 19.7).

The origin of these oscillations most likely lies in the nature of Earth's spin and its orbit around the Sun. Currently, Earth's axis is tilted some 23 degrees from a line perpendicular to the plane of its orbit around the Sun, and the orbit itself is slightly elliptical, or noncircular. The closest approach of Earth to the Sun happens to occur when the southern hemisphere is tilted toward the Sun, that is, during southern summer and northern winter. Because most of the continental mass lies in the northern hemisphere, this orbital state is one in which most of our planet's continental area does not experience the most summertime heating possible, because Earth is slightly farther from the Sun in July than in January. The difference in received sunlight, 8% from the closest to most distant point of the orbit, is significant in affecting climate.

Over a time of some 26,000 years, Earth's axis precesses around a fixed point in space, much like a toy gyroscope can be made to do by pushing its axis once it is set in motion. Viewed from the northern hemisphere, the current star closest to the north celestial pole, Polaris, will not forever be the north star: in 12,000 years Vega will be the pole star. The effect of this precession is to reorient the northern and southern hemisphere summers relative to the close and far points of the Earth in its orbit. Roughly 11,000 years from now, the northern hemisphere summer will occur when the Earth is closest to the Sun, opposite to the current state. Since plate tectonic motions are too slow to have shifted continental positions more than a few kilometers during that time, geography will be the same, and the heating of the northern hemisphere continental masses will be more severe 11,000 years from now than today.

Other oscillations in the motions of Earth are known to occur. The other planets of the solar system, in exerting very slight tugs on Earth, not only cause the axial precession but also slightly alter the magnitude of the tilt (from 21 degrees to 25 degrees) on 41,000-year cycles. The ellipse that is Earth's orbit drifts or rotates as well, which shortens the precessional period from

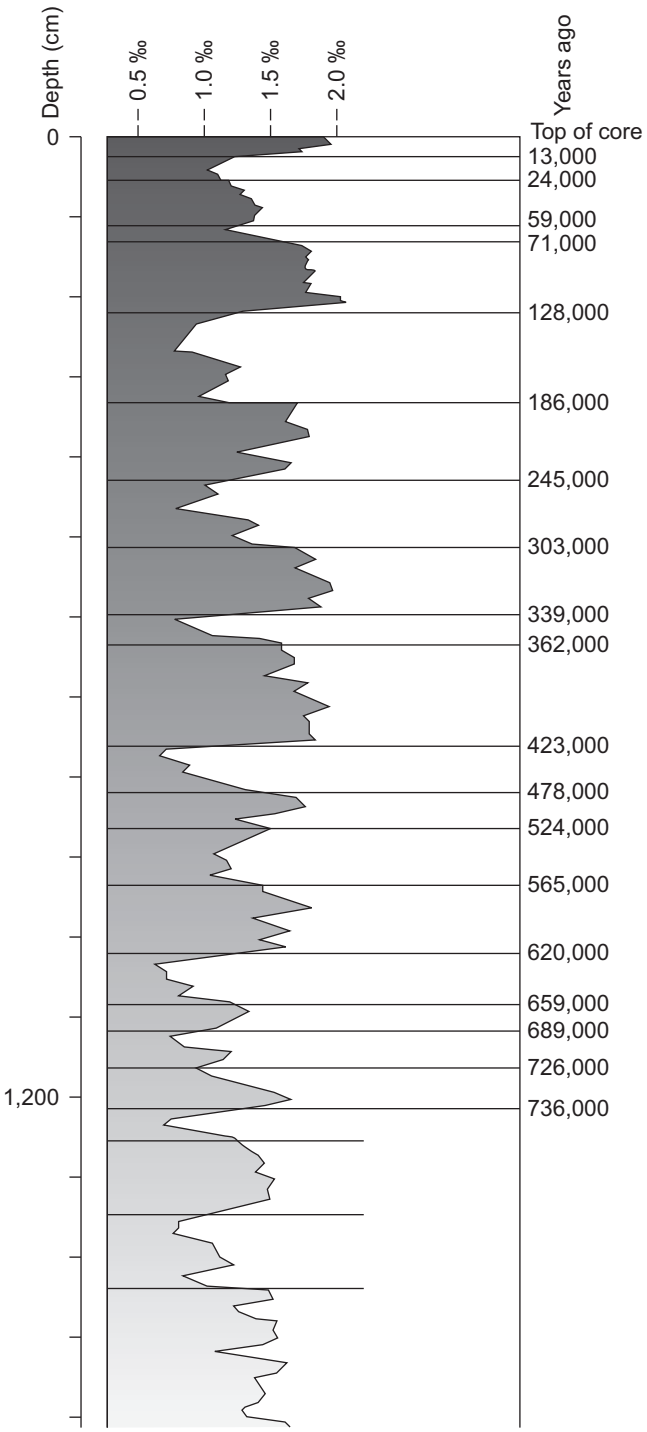


Figure 19.7 Data from a deep-sea drill core showing $^{18}O/^{16}O$ ratios over the past million years. The ^{18}O to ^{16}O value is an indirect measure of temperature: higher ^{18}O values (toward the left of the graph) mean more of Earth's water is locked up in continental glaciers, lower means more water is in the oceans (Chapter 6). On this graph, then, values toward the left indicate cold times, and toward the right warm times. The dates on the right are assigned to particular depths in the core sample by examining the magnetic orientation of sediment grains layer by layer and comparing this to the known pattern of reversals of Earth's magnetic field from a million years ago to the present. For further details on magnetic reversals, see Chapter 9. Adapted from Stringer and Gamble (1993).

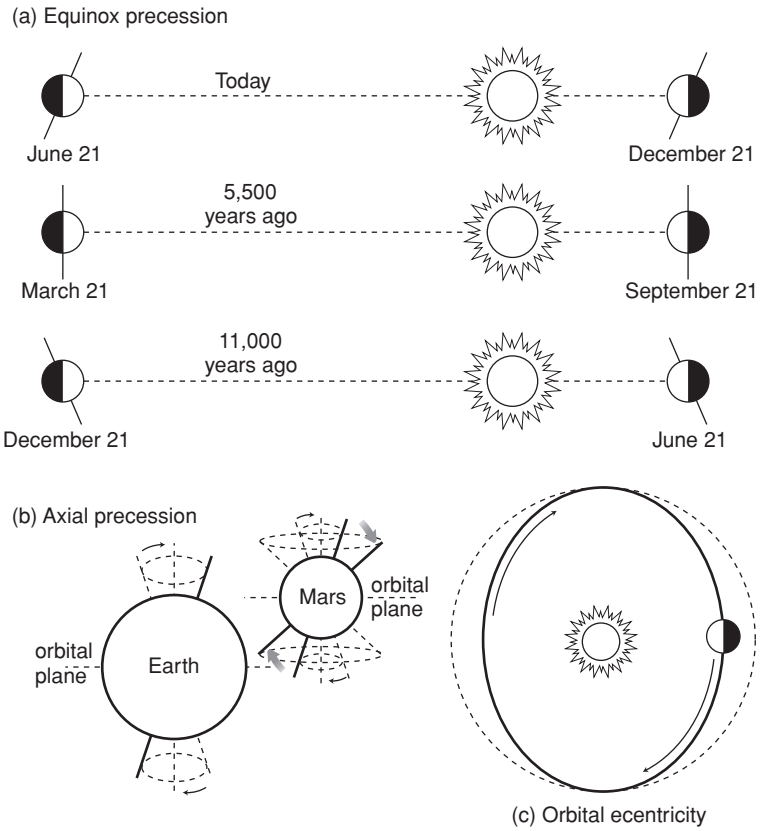


Figure 19.8 (a) Variation in the alignment of Earth's tilt and its orbit, causing a phase shift in the seasons. About 11,000 years ago, summer came to the northern hemisphere at a time when Earth was at perihelion; today northern summer occurs at aphelion. (b) Swings in the axial tilts of Earth and Mars over time; note that Mars suffers more extreme swings because it has no large moon to dampen the pull of the other planets. (c) Change in the shape of Earth's orbit on 100,000- and 450,000-year cycles and rotation or drift of the ellipse (indicated by the arrows).

26,000 to 22,000 years. Furthermore, these planetary tidal pulls also modulate the eccentricity of Earth's orbit in a complicated way that approximates two cycles at 100,000- and 450,000-year periods (Figure 19.8). The idea that all these cycles might affect climate goes back to the Scottish geologist J. Croll in the mid-nineteenth century. Serbian physicist M. Milankovitch, in the first part of the twentieth century, showed how these cycles affect the annual distribution of solar energy received by the Earth and hence the climate.

The relationship between the changing pattern of solar heating and glaciation is illustrated by the comparison in Figure 19.9. The lower graph shows the amount of sunlight received at 65° north latitude in July; this is one measure of the effect of the shifting orbital and tilt parameters on the distribution of sunlight falling on Earth. The upper chart, keyed to the same timescale, shows oxygen isotope data as a proxy for sea level and hence temperature (cold temperatures, lower sea level, higher ^{18}O). Careful examination shows that major decreases in the extent of glaciation occur when northern summer has its greatest solar heating.

The correspondence is not perfect but other effects come into play as well. In particular, the change in climate alters the carbon-cycle balance and hence the carbon dioxide abundance. Detailed models that account for the carbon dioxide effect show a much better correlation, especially at later times. The progression of glaciation episodes from 40,000 to 100,000 years as the Pleistocene progresses is not fully accounted for by orbital cycles; other effects not tied to the orbits (modulations in volcanic activity, minor plate movements changing ocean currents) might be involved. We defer a more detailed discussion of these effects to Chapters 21 and 22, in the contexts of the most recent glacial episode and concerns about the future of our planet's climate.

Why have the orbital cycles of the Earth amplified climate changes only in the last few million years? Earth's climate up to a few million years ago was sufficiently warm that the orbital changes in the pattern of sunlight on Earth (that is, in the strength and timing of the seasonal variations) were not enough to trigger formation of sufficient ice at high latitudes to force the onset of widespread glacial episodes. As global average temperature continued to decline in step with the removal of carbon dioxide, the amount of high-latitude ice formed during colder times in the Milankovitch cycles increased. Eventually, a point was reached several million years ago at which the amount of ice formed in the colder part of the cycle was sufficient to drive further cooling, and hence build up of massive ice sheets across large areas of the planet's continental masses. In effect, the gradually cooling conditions brought Earth's climate to the threshold of instability, and the colder portions of the orbital cycles forced the climate across that brink into glacial episodes.

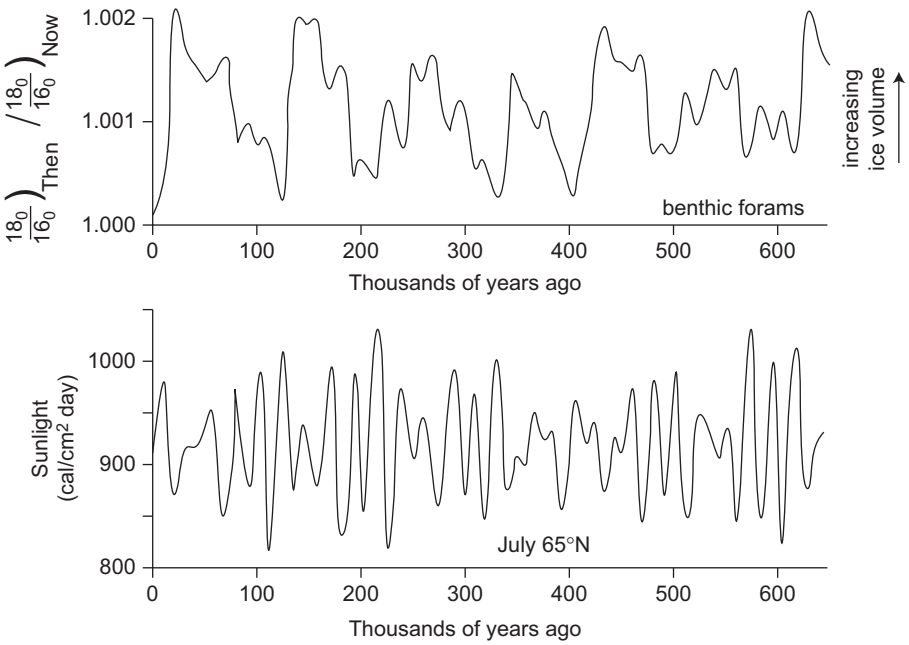


Figure 19.9 An examination of the effect of Earth's orbital and axial tilt variations on climate. Lower graph shows the amount of solar energy falling on the 65° latitude belt on Earth in July. This is affected both by tilt (more sunlight is received when the Sun is more directly overhead, i.e., when the northern hemisphere is tipped toward the Sun) and the varying distance from the Sun. The interplay between the various orbital and tilt cycles yields a complex pattern, much as musical instruments produce overtones and beats to achieve their rich spectrum of sound. The upper graph is a measure of the global temperature from oxygen isotope data. The climate is cooler when the relative oxygen-18 abundance is higher; on this graph the ratio $^{18}O/^{16}O$ is plotted *relative* to the present-day value (a ratio of ratios!). Reproduced with permission from Broecker (1985).

We do not know whether the Pleistocene ice age epoch is unique in its oscillatory behavior. Earth's orbit and spin likely have undergone cyclical variations throughout the solar system's history. Oscillations in ancient ice ages caused by such cycles may show only faintly in the geologic record, which becomes more imprecise with age. In the longest glacial period, some 340 million to 250 million years ago, there is evidence in sediments of cyclical changes in sea level, and some attempts have been made to ascribe these to climate oscillations induced by orbital and tilt cycles. Alternatively, since at least some of the more ancient ice ages were "deeper" in the sense that conditions were such as to plunge Earth into a climate much colder than that of the Pleistocene, the climate during those episodes would have been much less sensitive to orbital fluctuations. Indeed, the climate oscillations of the last two million years might require a fairly narrow range of glacial conditions close to the threshold of an ice-free climate. In this view, the environmental and biological effects of the severe climate oscillations between glacials and interglacials in the Pleistocene might be a very unusual feature of this last act of the Phanerozoic.

19.8 Saved from instability: Earth's versus Mars' orbital cycle

Dramatic as they are, Earth's orbital swings and associated climate changes are mild compared to those of Mars. Recent computer calculations by French and American scientists of the effect of other planets' gravitational pulls on Mars' tilt and orbit led to a surprising conclusion: the tilt of Mars behaves chaotically.

Chaos is a term co-opted by physicists to describe, not a state of complete randomness, but a common physical property of unpredictability in complex physical systems. Complex physical systems subjected to certain perturbations will not respond in a cyclical or deterministic fashion, but instead will break into a mode of unpredictable behavior or *deterministic chaos*. Certain characteristics of the behavior can be predicted by computer models, but the details of "what happens when" are lost.

For Mars, the chaos lies in the magnitude of the tilt of its axis. Rather than undergoing modest periodic swings in the amplitude of the tilt, the axis behaves chaotically and will swing from very small tilts of 10 degrees to as much as 50 degrees, on a wide range of different timescales. When and for how long a particular tilt excursion will occur cannot be predicted, but the amplitude range and general character of the swings can. Evidence for wide swings in the Martian rotational axis is present in a series of layered deposits thought to be dust and ice, residing near the poles (Figure 19.10). Alternating epochs of deposition of these two materials can be achieved through large swings in the tilt of the axis.

Disturbingly, similar computations for Earth show equally dramatic results, but evidence for wild swings is lacking in the geologic record. The stabilizing influence seems to be Earth's Moon. The tidal pull of the Moon on Earth is strong compared to the pull of the other planets (because of their greater distances), and the Moon's pull strongly damps out large excursions in Earth's tilt that might otherwise occur. However, had

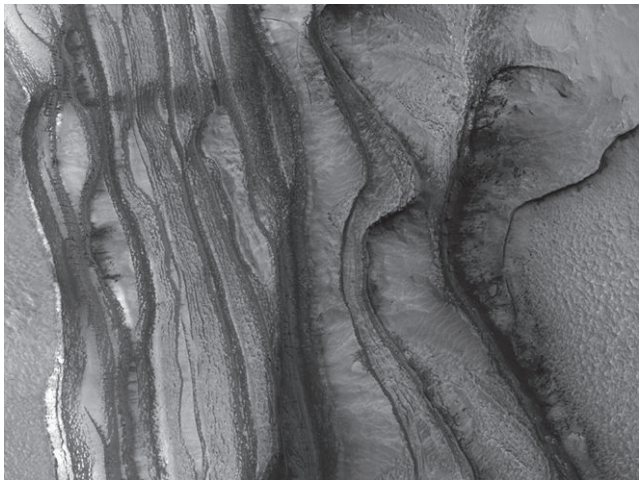


Figure 19.10 Layered deposits of dust and ice at the south pole of Mars. See color version in plates section.

the Moon been absent, a wild set of Mars-like swings was not inevitable for the Earth. The same models that predict the Martian oscillation predict that an Earth spinning twice as fast as our own would have a precessional period (the time for the precessing axis to make one cycle around the sky) much shorter than 26,000 years, and would be stable against variations in the axial tilt. An Earth with a twelve-hour day would be inoculated against wild swings, but for most of us 24 hours per day are not enough.

If indeed the presence of a large natural satellite is responsible for stabilizing Earth's tilt, and hence preventing frequent and drastic excursions of climate, the implications may be profound for life elsewhere. Finding an Earth-like planet around another star would not necessarily be enough to buoy the hopes of finding advanced life: one would need to ascertain the architecture of the planetary system, where the giant planets are located, whether the planet of interest has a large moon, whether the planet itself spins rapidly enough to obviate the need for a moon. Some or all of these parameters will be difficult, but not impossible, to ascertain from telescopic observations.

19.9 Effects of the Pleistocene ice age: a preview

With the onset of the oscillatory ice ages, the less stable climate contributed to species extinctions, extensive migrations, and the development of new species and even genera of animals. Of much interest to us is the coming of human-like creatures and then humans as a part of this 2-million-year time of change. In the next chapter, we explore one of the most startling results of the long evolution of this habitable planet: the coming of the age of humankind.

Summary

The Phanerozoic provides an excellent record of climate change right up to the present. On the longest timescales of hundreds of millions of years, the cycle of break up and reassembly of a single global continent – the result of plate motions – leads to changes in the patterns of ocean circulation, of abundance of high plateaus and hence continental glaciers, and of volcanism and erosion, which affect carbon dioxide levels. Ice-free times of great warmth, such as the Cretaceous, may reflect the presence of a single supercontinent that has existed for some time; with little continental material at the poles, and topography ground down by erosion, there is limited surface area for the ice that provides a positive feedback in cooling the Earth. Ocean currents are free to efficiently move heat from the warm equator to the poles in a vast superocean unimpeded by continental material. Break up of the supercontinent and dispersion of the fragments changes ocean circulation patterns, moves landmasses to high latitudes and through the re-collision of fragments raises large plateaus that can scrub CO_2 out of the atmosphere by forcing enhanced rainfall on regional scales. In the mid-Tertiary, the global climate began a cool-down that would see its climax in the last two million years of Earth history: the oscillations between glacial and interglacial climates. The underlying cause of the modulations is almost certainly variations in the orientation of the Earth's axial tilt relative to the perihelion of its orbit about the Sun, as well as periodic changes to the shape and orientation of the orbit itself. This change in distribution of sunlight amplifies a number of other effects such as ice cover and even CO_2 levels, to create the dramatic differences between the glacial and interglacial times. As dramatic as these are, Earth might have suffered even wilder swings had it not possessed a large Moon: our neighboring planet Mars has an axial tilt that dips back and forth, becoming as large as twice or more that of Earth, thanks to its exposure to the gravitational tugging of Jupiter.

Questions

1. Early models that attempted to simulate the onset of ice ages were simple and often done in one dimension (latitude). These models were unstable in the sense that adding a little ice would cause the entire Earth to freeze over, permanently. What key variable aspect of Earth's climate might be lacking in such models?
2. If the Cretaceous experienced such warm temperatures, might it have approached the threshold of a moist runaway greenhouse as discussed for Venus in Chapter 15? What is the key for such a runaway – temperature or solar flux?
3. Could a planet with a more highly elliptical orbit than Earth's recover from glacial swings in which the entire surface area becomes ice covered?
4. What other isotopic ratios might one use to detect the signature of ancient glaciations besides $^{13}C/^{12}C$? Considering the need to go back billions of years, what kinds of problems might arise in interpreting such isotopic data?

References

Barron, E. J. 1983. A warm, equable Cretaceous: the nature of the problem. *Earth-Science Reviews* **19**, 305–38.

Barron, E. J. 1992. Paleoclimatology. In *Understanding the Earth: A New Synthesis* (G. C. Brown, C. J. Hawkesworth, and R. C. L. Wilson, eds). Cambridge University Press, Cambridge, UK, pp. 485–505.

Barron, E., Fawcett, P. J., Peterson, W. H., Pollard, D., and Thompson, S. L. 1995. A "simulation" of mid-Cretaceous climate. *Paleoceanography* **10**, 953–62.

Broecker, W. 1985. *How to Build a Habitable Planet*. Eldigio Press, New York.

Cloud, P. 1988. *Oasis in Space: Earth History from the Beginning*. W. W. Norton, New York.

Crowly, T. J., Yip, K.-J. J., and Baum, S. K. 1993. Milankovitch cycles and carboniferous climate. *Geophysical Research Letters* **20**, 1175–8.

Jouzel, J. and 31 others. 2007. Orbital and millennial climate variability over the past 800,000 years. *Science* **317**, 793–6.

Marshall, H. G., Walker, J. C. G., and Kuhn, W. R. 1988. Long-term climate change and the geochemical cycle of carbon. *Journal of Geophysical Research* **93**, 791–801.

McGoweran, B. 1990. Fifty million years ago. *American Scentist* **78**(1), 30–9.

Meert, J. G. and van der Voo, R. 1994. The Neoproterozoic (1000–540 Ma) glacial intervals: no more snowball Earth? *Earth and Planetary Science Letters* **123**, 1–13.

Milne, D., Raup, D., Billingham, J., Niklaus, K., and Padian, K. (eds) 1985. *The Evolution of Complex and Higher Organisms*. NASA SP-478. US Government Printing Office, Washington, DC.

Murphy, J. B. and Nance, R. D. 1992. Mountain belts and the supercontinent cycle. *Scientific American* **266**(4), 84–91.

Pälike, H. and Hilgen, F. 2008. Rock clock synchronization. *Nature Geoscience* **1**, 282.

Peixoto, J. P. and Oort, A. H. 1992. *Physics of Climate*. AIP Press, New York.

Pierrehumbert, R. T. 2005. Climate dynamics of a hard snowball Earth. *Journal of Geophysics Research* **110**:D01111.

Pierrehumbert, R. T., Abbot, D. S., Voight, A. and Knoll, D. 2011. Climate of the neoproterozoic. *Annual Reviews of Earth and Planetary Science* **39**, 417–60.

Raymo, M. E., Ruddiman, W. F., and Froelich, P. N. 1988. Influence of late Cenozoic mountain building on ocean geochemical cycles. *Geology* **16**, 649–53.

Rinaldo, A., Dietrich, W. E., Rigon, R., Vogel, G. K., and Rodriquez-Iturbo, I. 1995. Geomorphological signatures of varying climate. *Nature* **374**, 632–5.

Rohde, R., Curry, J., Groom, D. *et al.* 2011. Berkeley Earth temperature averaging process. http://berkeleyearth.org/pdf/berkeley-earth-averaging-process.pdf.

Shackleton, N. J. and Opdyke, N. D. 1973. Oxygen isotope and paleomagnetic stratigraphy of equatorial Pacific core V28-238. *Quaternary Research* **3**, 39–55.

Stringer, C. and Gamble, C. 1993. *In Search of the Neanderthals: Solving the Puzzle of Human Origins*. Thames and Hudson, London.

Tattersall, I., Delson, E., and Van Couvering, J. 1988. *Encyclopedia of Human Evolution and Prehistory*. Garland Publishing, New York.

20

Toward the age of humankind

Introduction

Earth's evolutionary divergence from the neighboring planets of the solar system, beginning with the stabilization of liquid water, culminates in the appearance of sentient organisms sometime within the past 1 million to 2 million years. The fossil record is abundant in its yield of creatures intermediate in form and function between the great apes and modern humans; new discoveries seem to be made with increasing pace. But hidden between and among the fossil finds are the details of how and why we came to be. Even as we acknowledge our common origins with the life around us, the singular results of sentience – art, writing, technology, civilization – are surprising and enigmatic.

The story of human origins is not simple, and changes with every new fossil find. Therefore, this chapter attempts only a sketch of the evidence and the lines of thought current in today's anthropological research. It begins with a broad view of the climatological stage on which these events took place. It ends with a focus on the closing act of human evolution, the coexistence of modern humans with a similar but separate sentient species in Europe and the Middle East – the Neanderthals.

20.1 Pleistocene setting

The earliest fossils along the lineage toward humanity exist in the Pliocene epoch, prior to the Pleistocene, during a time of relative climate stability. The pace of human evolution picks up in the Pleistocene, and species close enough in form to us to warrant assignment to the genus *Homo* (Latin, man in the sense of humans) appear close to, but perhaps slightly before, the time when climate shifted into an ice-age pattern of glacial and interglacial episodes.

The effect of glaciers was profound. During the depths of the glacial episodes, ice sheets stretched across significant parts of North America, Asia, and Europe. These sheets exceeded 3,000 meters in thickness in places, and hence acted like huge mountain ranges in diverting air flow and weather patterns by thousands of kilometers. Ocean currents were affected by changes in the amount of sea ice year round, by alterations in salt content, and by the patterns of rainfall and snowfall. The rise and fall of sea level by more than 100 meters opened and closed overland routes between continents. The amount of plate movement of continents was relatively small, no more than tens of kilometers over a million years (Chapter 9), but this was more than made up for by the oscillations associated with the advance and retreat of glaciers. Such oscillatory effects acted to move ecological niches significantly on timescales ranging from 100,000 to 10,000 years, and probably even less. Forests waxed and waned over large areas; food supplies changed dramatically between cold–dry and warm–wet episodes. Animal species encountering such changes either perished or migrated vast distances, and many opportunities for speciation (formation of new species) must have been available as small groups became isolated (Chapter 18).

The foment caused by the instability of climate is reflected in the extinction of a number of mammalian species during this time. It also may have served as the stimulus for a dramatic change in the kinds of primate species present in Africa and possibly Asia. The alternate waxing and waning of savanna versus forestland, so different in the kinds of species and survival styles they support, may have been at the nexus of the production of new primate lineages and extinction of the old.

20.2 The vagaries of understanding human origins

The fossil record of human origins has become remarkably rich, and the ability to do forensic analyses – even DNA analysis in the case of Neanderthals – has provided a wealth of information on the history of the human species and its precursors.

Nonetheless, as discussed in Chapter 8, the vast majority of living organisms are broken down after death without their body forms being preserved. The very few that die in environments resulting in fossil production must serve as the faint signposts of an evolutionary process involving vastly larger numbers of organisms. Therefore, the story of human origins will always remain incomplete.

With human evolution, this problem of incompleteness is compounded by another challenge, what might be called the "goldfish bowl" effect. Human origins means *our* origins and, as such, any discoveries are subjected to intense scrutiny by the public. There is a natural tendency, with any announced new fossil find, to hope that it solves "the" puzzle, so that often unjustified conclusions are drawn by the press, as well as by anthropologists themselves. Adding to the emotional foment are the personal religious beliefs held by individuals; for some religions the notion of an animal origin for human beings, without supernatural intervention, is heretical and offensive.

For these reasons the history of the search for physical evidence of human origins has been replete with dramas played out in social and cultural arenas, beginning even before publication of Darwin's ideas on human origins in his 1871 book *The Descent of Man*. The notorious Piltdown hoax of 1913, a fabricated skull constructed essentally of an ape jaw and human cranium, may have been an interesting scientific Rorschach test but also created a credibility gap with long-term repercussions. The "Scopes Monkey Trial" of 1925 was a famous legal challenge to a Tennessee law restricting the teaching of evolution; it centered on the conflict between Biblical scripture and biological understanding of species origins. Remarkably and regrettably, dramas akin to the Scopes trial are played out in US school boards and on the campaign trail almost *ad nauseum*. But it must be remembered that philosophers have long reflected on the status of humans as a type of animal; Aristotle called us the "rational animal."

20.3 Humanity's taxonomy

To appreciate the search for human origins requires returning briefly to the discussion of taxonomy of Chapter 18. All human beings alive on Earth are members of the same species, *Homo sapiens* (Latin, wise man), in turn the sole representative of the genus *Homo*, which in the past has contained a number of other species. We are members of the family *Hominidae*, comprising several now-extinct genera, along with *Homo*, chimpanzees, and gorillas. The inclusion of the African great apes and humans in the same family is the recent resolution of a long-standing taxonomic argument; previous classifications putting apes in a separate family were flawed because physiologically (and genetically) humans are more closely related to chimps and gorillas than any of the three are to the orangutan.

The apparently large gap between ourselves and nearest animal relatives arises in part because many other creatures classifiable in the genus *Homo* are extinct. Whether by climate change or competition from our most successful immediate ancestors, we sit out on a rather isolated limb of the primate family tree.

In what follows, we briefly sketch a picture of human evolution based on key fossil species identified to date, one that is summarized in Figure 20.1. As in any such narrative, the simplicity of the results belies the decades of controversy, discovery, and revision that have preceded and will follow this particular moment in anthropology. Consider that you have been given the task of assembling a jigsaw puzzle. You do not know what the final picture will look like, nor do you know how many pieces there are. The pieces are not in a box; they've been scattered around town and you must find them. Some are in such poor condition that their edges are frayed, torn, or missing; nonetheless you must find the pieces and, through trial and error, assemble the final image. Such is the essence of the anthropological search for how humankind came to be.

20.4 The first steps: Australopithecines

Africa seems to be the source of the most ancient fossils in our ancestral family tree. This continent is rich today in primate species and, particularly in the equatorial regions, would have exhibited relatively gentle environmental fluctuations in response to the overall climate instability of the Pleistocene and preceding Pliocene. Much confusion and uncertainty about whether Africa or Asia was the origin point for the outward radiation of new hominid species seems, for the moment, to be resolved in favor of Africa.

Studies of genes in apes and humans, coupled with estimates of the rate of mutation of such genes (Chapter 12), lead to the conclusion that the African apes (chimps and gorillas) and humans had a common and now-extinct ancestor as recently as 5 million years ago, but no earlier than 9 million years ago. Indeed chimp and human genomes are approximately 95% identical. Therefore, understanding what happened in the split and "who was there" in the fossil record on each side after the split is difficult, but paleontologists now generally agree that the genetically estimated timescale is probably right. Around the time of the split, there existed two species in the genus *Ardipithecus* (Latin for "chimp-like"), present in the form of jaw and cranial fragments, bearing the signature of the great apes but differing in detail from gorillas, chimpanzees, and ourselves.

Beyond this point, a variety of species in two different genera (plural for genus) begin to appear in the fossil record, principally in Africa. Over 300 specimens define *Australopithecus* ("southern ape") *afarensis*, a genus that lived in Africa over a span of time from 2.5 million to 3.4 million years ago, and perhaps longer.

Still very much ape-like, with little to indicate a direction toward human ancestry, *A. afarensis* is distinguished by the large number of specimens, its broad span of time, and its representation in a fairly complete skeleton known popularly as Lucy. Its younger age than the genetically determined split of human from the great apes, a demonstrably upright posture, and a larger brain size than the chimpanzee suggest that it is an early species on the road to humankind. Nonetheless, were we to see a living *afarensis* today, it would seem to us no more than a fascinating ape that happened to walk upright and was somewhat smarter than a chimpanzee. Other species of the genus *Australopithecus* existed down to about 1.5 million to 2 million years ago. A separate (perhaps offshoot) genus, *Paranthropus* ("near man"), also is represented in this time by several species,

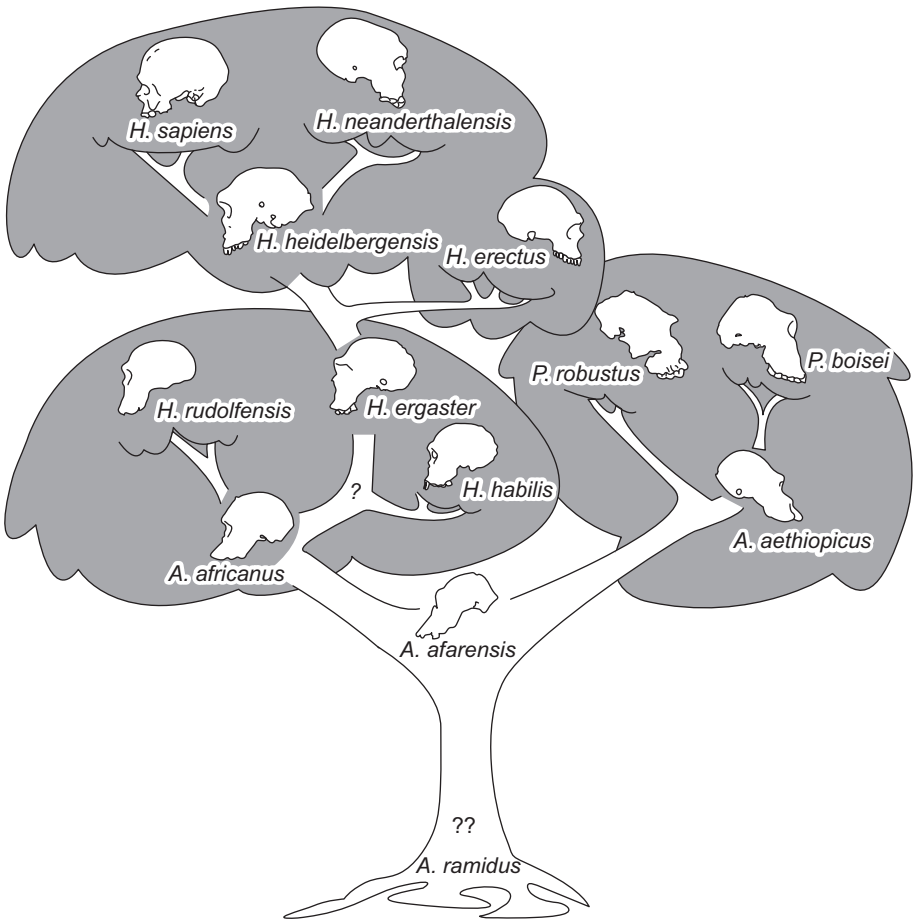


Figure 20.1 Species related to, and in some cases ancestral to, modern humans, assembled in a notional genealogy. Key to genus names: *A.* = *Australopithecus*, *P.* = *Paranthropus*, *H.* = *Homo*, our genus.

and extends to a million years before present, well into the time of the genus *homo*. *Paranthropus* was more robust than either *Australopithecus*, or *Homo*, had a more restricted diet, and is for all intents and purposes another ape, destined for extinction. The three overlapping genera over a period of almost 2 million years made the African continent a far richer tapestry of hominid species than is all of today's world combined.

20.5 The genus *Homo*: Out of Africa I

Between 3 million and 2.4 million years ago, the African climate shifted to a dryer, cooler regime than had dominated previously, and the first species of our genus – *Homo* – then appeared. Whether the changing climate stimulated contemporaneous dramatic changes in the Hominidae line is unclear. An old picture is that the human lineage resulted from creatures who moved out from the forests into the plains, leaving behind the lineage that became great apes. This view is now held in very low regard, based on evidence that both *Australopithecus* and *Paranthropus* were adapted to partially open, woodland conditions. But certainly fluctuating environmental conditions caused shifts in the extent and nature of woodlands, shifts that provided a greater opportunity for isolation of groups, followed by speciation encouraged by environmental stresses – and extinction of those that could not adapt.

Between 2.5 million and 2 million years ago, several different species appear in Africa that were too human in appearance and sophisticated in behavior to merit inclusion in the genus *Australopithecus*; instead, they are the earliest members of the genus *Homo*. They possessed crania larger and differently shaped than *Australopithecus*. They appeared to fashion crude stone tools to assist their hunting and food preparation. The most successful member of the genus *Homo* in terms of species longevity, *Homo erectus* (upright man), appears around 2 million years ago or a bit later. *Erectus* had a larger cranial capacity and more human features than the *Homo* species before it. There is evidence for more extensive stone modification and use as tools. *Erectus* as a species is recognizable for a million years, the longest lived member of the genus *Homo* to date.

Only shortly after the appearance of *Homo* in Africa, members of this genus began migrations eastward into Asia. Recent finds of *Homo erectus* in eastern Asia that have ages approaching 2 million years suggest a prompt dispersal in that direction. Migrations of *Homo erectus* populations would continue for over a million years, eventually leading to the establishment of groups in Europe as well (with a continuous lineage that extends almost, but not quite, to the present). Hypotheses as to the origin of this propensity for travel include the changing

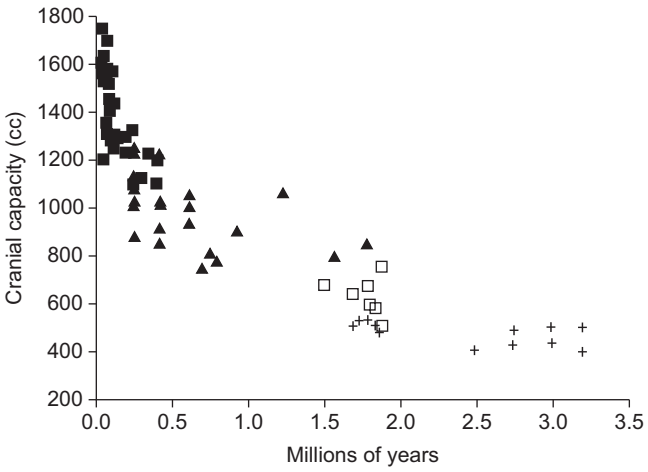


Figure 20.2 Cranial size in hominid species as a function of time, adapted from Mellars (1996). The units of volume of the cranium are cubic centimeters.

climate, driving many species toward dispersal or extinction, and the tendency, suggested in fossil remains, of African *Homo* to range widely in its scavenging and hunting forays. Whatever the cause, the wandering nature of *Homo* distinguished it from its predecessors.

It is with the Out of Africa I migration that the story of human evolution takes a complex turn. Because *erectus* and similar *Homo* species had spread onto three continents (Africa, Asia, and Europe), the geographical area covered was too large to permit gene transfer by interbreeding among groups. Instead, the fate of the various *Homo* groups became decoupled from one another, and a complex and poorly understood pattern of emergence of various post-*erectus* species is played out over many hundreds of thousands of years. The situation, by 200,000 years ago, was the apparent existence of post-*erectus* species on three continents, with brain sizes approaching or equaling present-day values (Figure 20.2), and whom, for want of a better term, are called "archaics." The pace of change had accelerated, perhaps because of increased climate fluctuations, the propensity for migration that would naturally produce isolated populations ripe for further speciation, or other causes. That situation persisted up to nearly the present day – but in a blink compared to geologic time, all such species disappeared except our own.

20.6 Out of Africa II

As in all sciences, controversy rages in anthropology over crucial parts of the story of human origins. Two views exist as to what happened to effect a transition from the post-*erectus* populations scattered across Europe, Asia, and Africa to the present situation of a single, modern species, *Homo sapiens*, occupying all the Earth (Figure 20.3).

The *multiregional* origin posits that the post-*erectus* populations encountered each other enough to allow interbreeding to maintain a single, archaic-human species, but not enough to erase regional differences. This species evolved separately and

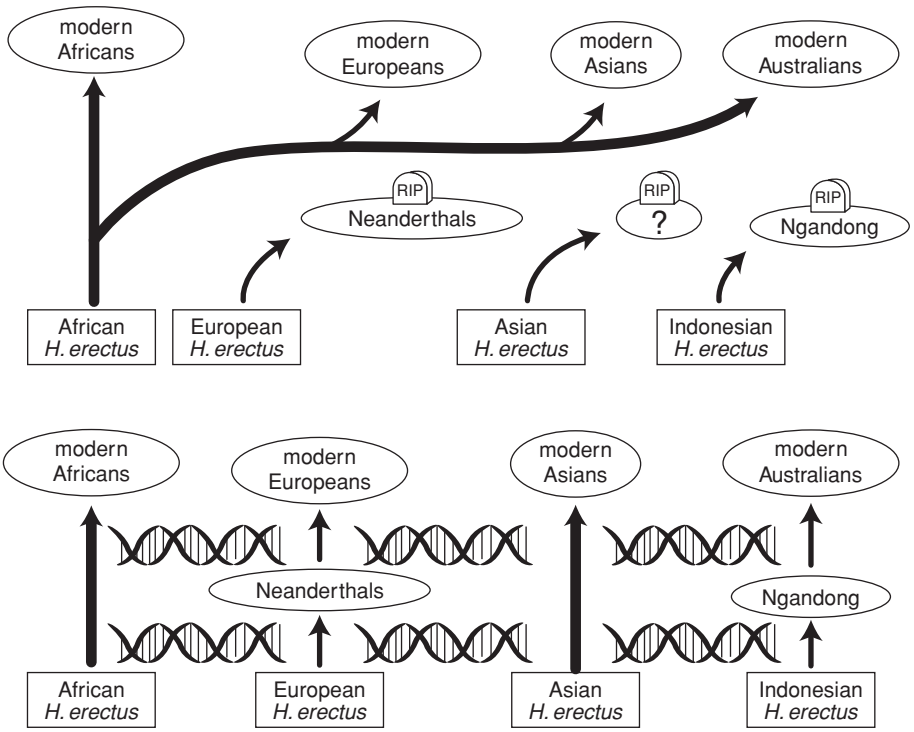


Figure 20.3 Schematic comparison of the replacement (top) and multiregional (bottom) hypotheses for the origin of modern humans. The double helixes in the multiregional model symbolize gene transfer by occasional interbreeding. Time flows upward in each model. Modified from original figure by Christopher Stringer from Stringer and Gamble (1993) by permission of Thames and Hudson.

semi-independently on the three continents into modern *Homo sapiens*, with the genetic transfer rate sufficient to ensure maintenance of a single species. The origin of races, in this view, is very ancient. The alternative view, *replacement*, posits that a final, late speciation event occurred somewhere in the world, and the new species, modern *Homo sapiens*, spread outward from that point and supplanted the existing archaic peoples on all three continents. Under this hypothesis, the establishment of racial groups is a very recent phenomenon associated with modern *Homo sapiens*' adaptation to regional climate variations.

The multiregional hypothesis relies largely on regional differences in appearance among *erectus* and later archaic specimens that are vaguely similar to those among the various races of modern humans today. However, the hypothesis suffers from very serious drawbacks. It relies on an excessively delicate balance of interbreeding: enough to ensure that humankind did not diverge into separate species (while evolving substantially from *erectus* into *sapiens*), yet not so much that regional differences were erased. It is not consistent with the extreme commonality of the genome that we as modern humans possess.

The replacement model, on the other hand, has support from the genomic record and the fossil record. Genomic analysis of our origins relies on genetic variations among present-day human populations to trace backward the region and time in which a putative single speciation event occurred. The challenge with this approach, of course, is that every child is the product of the shuffling of genes from a father and a mother. However, genes contained in the cellular mitochondria (Chapter 12) are inherited almost always from the mother alone. Provided, then, that there has been a continuous lineage from an initial speciating group to the present, examination of mitochondrial DNA differences might yield the origin location and time.

This technique, pioneered by University of California researchers A. Wilson and R. Cann, required extracting mitochondrial DNA samples from people all over the world. The results suggested two robust conclusions. First, all human mitochondrial DNA is extremely similar, indicating that very little time has elapsed in which racial differences have built up. Second, African peoples show relatively more variation in their mitochondrial DNA than do people from other continents, suggesting the Africans have had somewhat more time to build up such variations – that is, modern humans have existed in Africa the longest. By using known divergence rates among modern groups, it is possible to derive a timescale for when modern humans first arose. The result of the analysis is that all humans alive today arose from a very small, interbreeding, group in Africa some 100,000 to 200,000 years ago.

Evidence in the fossil record is consistent: it shows *Homo sapiens* with modern physiology and possessing extensive advanced tool kits, appearing in southern Africa and East Africa some 120,000 to 135,000 years ago, and later in Northern Africa. The oldest Asian samples of modern *Homo sapiens* seem somewhat younger – consistent with a migration from Africa. Modern humans appear in the Middle East by about 90,000 years before present, and Europe perhaps 45,000 years ago. Once spread into these many lands, with their different climates and physical demands, flowered into the many races of today – but all members not just of one species, but of one subspecies *Homo sapiens sapiens*.

When taken in total, the fossil and genetic evidence favor a model in which modern *Homo sapiens* first speciated somewhere in Africa, between 100,000 and 200,000 years ago, and quickly spread via migrations throughout the world. What happened in its encounters with the archaic humans already present in Africa, Asia, and Europe was a long story, spanning many tens of thousands of years. There is little in the way of physical clues to reveal the nature of the replacement process. The best documented – and perhaps longest – overlap of modern and archaic humans was in Europe and the Middle East, home of the archaic Neanderthals. It is worth focusing on this last act of human evolution, not merely for its own sake, but as a cautionary tale as we contemplate how our own species might interact or coexist with intelligent life elsewhere in the cosmos.

20.7 Final act: Neanderthals and an encounter with our humanity

All history contains lessons about our own humanity, but these are often couched in ambiguous terms. The same is true for the saga of the Neanderthals in Europe and the Middle East (Figure 20.4), over a time (perhaps between 30,000 and 100,000 years) when this interesting species crossed paths with anatomically modern humans.

20.7.1 Climate setting

The first hints of characteristic Neanderthal features and tools extends back 300,000 years ago or so, in Europe and western Asia. The last physical evidence for Neanderthals is found in southern Spain as recently as 27,000 years ago. This long interval spans two major glacial episodes of the Pleistocene, as seen in the oxygen-isotopic record from sea sediment cores, displayed in Figure 20.5. Separating the two is Earth's penultimate interglacial (we are living in the most recent one), extending from 118,000 to 126,000 years ago. During the interglacial time, forests would have covered large areas of continents, including Europe and Asia. A variety of large mammals seen today in restricted regions were present throughout the world; hippopotami and elephants were found in the British Isles, for example.

More recently than 118,000 years ago, temperatures initiate a somewhat bouncy descent toward the most recent glacial, though they do not begin to approach the cold of the previous glacial until 70,000 to 80,000 years ago. In the range of the Neanderthals, broken woodlands existed, probably supporting herds of grazing animals; many of the large mammals disappeared as the glaciation intensified.

From 70,000 to 30,000 years ago, ice sheets advanced and retreated in rapid oscillations. Major cold episodes (such as 60,000 to 70,000 years ago) must have accelerated extinctions of a variety of animals; forests retreated southward and previous woodlands likely became open tundra in Europe. In the Middle East, periods of clement climate may have existed at these colder times, encouraging migration of Neanderthals down to that geographic crossroad of the world. The final mild climate

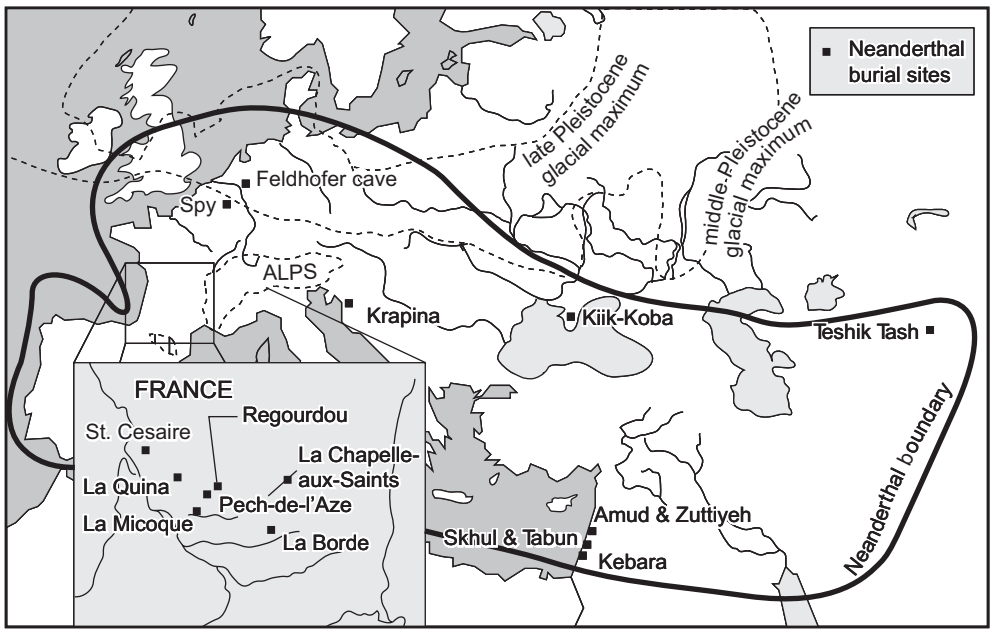


Figure 20.4 General geographic areas occupied by Neanderthals. Dashed lines indicate the extent of the glaciers during the middle and late Pleistocene. Modified from original figure by Annick Peterson from Stringer and Gamble (1993) by permission of Thames and Hudson.

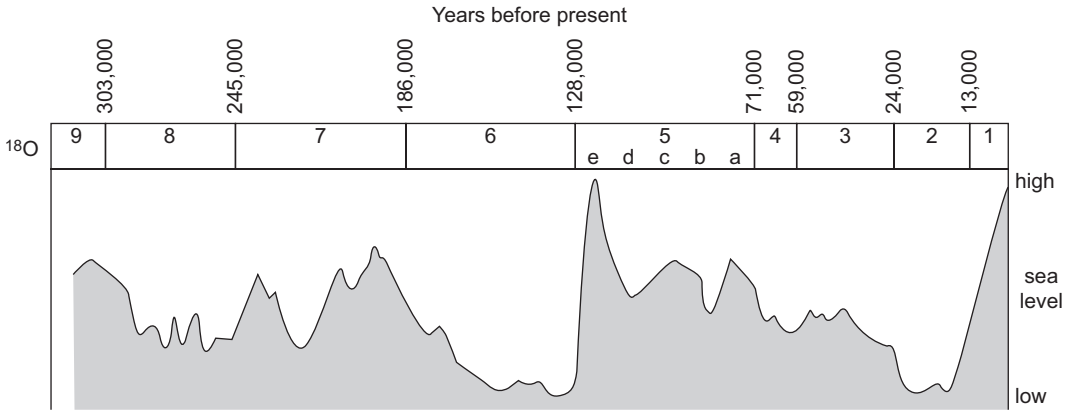


Figure 20.5 Sea level, and hence temperature, over the past 300,000 years from oxygen-18 data in seafloor sediments (see Chapter 6 for discussion of technique). Times of high sea level, hence less ice, are warm; low sea level indicates colder, glacial epochs. The numbers from 1 to 9 are standard labels for glacial and interglacial episodes. From Stringer and Gamble (1993) by permission of Thames and Hudson.

episode before the peak of the last glaciation occurred 40,000 years ago. The climactic freeze was reached about 19,000 years ago, after the extinction of the Neanderthals.

Early Neanderthals, with somewhat different features than their "late" Neanderthal descendants, existed from perhaps 300,000 to 130,000 years ago. The time after that, up to perhaps 40,000 years ago, was really the heyday of the Neanderthals, with characteristic stone cultures and stable physical features. During this time, Neanderthals made incursions into the Middle East, interleaving with peoples of more modern appearance and stone cultures. From 40,000 years to their extinction, the Neanderthal populations declined in geographic distribution, while innovating through imitation of stone tool types brought by the modern peoples emigrating to Europe.

20.7.2 Physical features of Neanderthals

Neanderthals were not the stooped over, ape-like, brutish cousins of the depictions of the popular literature. They were short but very robust people, with broader and deeper muscle attachments in their bones, and hence more massive musculature, than the average for any modern populations. Although their hip attachments differ from ours in encouraging more stress on the sides of their thighs than on the front and back (making easier the squatting and sideways movements typical of foraging activity), the fossil remains of their skeletons are consistent with fully upright postures.

It is the head of Neanderthal people that most dramatically outlines the difference from all modern humans (Figure 20.6).

Figure 20.6 The author as (left) *Homo neanderthalensis*; (right) *Homo sapiens*.

The Neanderthal skull has very heavily enlarged brow ridges; a cranial vault that is low and somewhat flattened relative to that of modern humans; more massive jaws and teeth relative to the rest of the skull than in modern humans; a huge, broad nose, and virtually no chin. However, the cranial capacity of Neanderthals equaled or exceeded that of modern humans. To accommodate the brain in the more flattened skull, Neanderthal heads had a more prominent rear "bun," than do modern humans. Human skulls are constructed such that they grow outward as the brain grows during infancy and childhood. Presumably Neanderthal's did the same, hence the shape of the skull, which would strike any human being today as being very odd, reflected a differently shaped (and presumably differently functioning) brain, but one on average somewhat larger than ours.

Many of the features of human and Neanderthal heads likely related to the need to support chewing and grinding forces. Our skulls have high front domes, providing adequate support against muscular forces; Neanderthal brow ridges did the same in the absence of the high forehead. Our chins likewise provide structural support during chewing, and are a somewhat unusual innovation in the hominid line; Neanderthal jaw stresses were supported by more traditional heavy bones.

The striking stockiness of Neanderthal bodies (both male and female) and evidence for large muscles could readily be argued as a result of a more strenuous physical lifestyle. However, such features are present in preadolescent Neanderthal children, whose ages at death are easily dated from the state of their dentition. The stocky build surely enabled a physically demanding lifestyle, but may have had its origin in adaptation to the very cold climate that characterized Europe during much of the Neanderthal heyday. This is the case among modern people who live in very cold climates – but not nearly as extreme as that of the Neanderthals.

A further clue to this adaptation lay in the heroically sized nose. Anthropologists have argued that it could serve two possible (and likely simultaneous) functions: warm the frigidly glacial air as it is inhaled, and allow for a greater volume of inhalation with a consequently higher tolerance for physical exertion. Enthusiasts for backcountry winter sports know the hazards of overexertion and consequent sweating: hypothermia (a loss of body temperature control) and death can result. A bigger nose is an adaptation allowing a high-exertion lifestyle in the cold.

Having emphasized the differences from modern humans, it is now necessary to remark upon how close the Neanderthals are to us in their appearance. Meet one in modern dress in an office and you would do a double-take: this seems to be a human being, but what a strange head and face! More different than any of the remarkable variety we share as modern humans, one of the great enigmas of the Neanderthals is the juxtaposition of the oddness with the closeness to modern humans. Most anthropologists today hold the view that Neanderthals are *Homo neanderthalensis*, a different species sharing the same genus as modern humans. The physiological differences between Neanderthals and modern humans are larger than between other primates that are, without controversy, classified as different species.

The origin of Neanderthals seems to lie in pre-existing populations of *Homo erectus* or a successor species *Homo heidelbergensis*, resident in Europe as well as western Asia for many hundreds of thousands of years. Many of the traits of Neanderthal features can be seen in fossils from prior to 300,000 years ago in England, Germany, Greece, and France – remains that seem transitional between *erectus/heidelbergensis* and Neanderthal. Far removed from the changes occurring in Africa that led to modern *Homo sapiens*, the Neanderthals were an evolutionary event in and of themselves – a distinct population of *Homo* evolved from ancestors who migrated out of Africa or Asia long before the speciation event that produced modern *Homo sapiens.*

Neanderthal fossil remains show differences from individual to individual. However, these differences are smaller than are the differences between individual members of today's modern humans. Our species has spanned the globe, adapting to a range of climates far greater than those the Neanderthals contended with. It is not surprising, then, that we should be a more varied species than Neanderthal. Equally important is the lack of transitions between Neanderthals and modern humans. With only a few controversial exceptions from the Middle East, the fossil record seems to be telling us that there is no transitional form, no people that reflect a strong heritage of interbreeding between coexisting Neanderthals and modern (or near-modern) humans who lived at the same time.

Beginning in 1997, extraction and analysis of DNA samples from Neanderthal bones has been possible. In 2010, scientists announced that the Neanderthal genome had been sequenced. The Neanderthal genome is about the same size as the human genome, and is identical to ours to a level of 99.7% (this is comparing the ordering of the lettering in the nucleotide bases). Using an average rate of mutations, Neanderthal and human lineages diverged between 270,000 and 440,000 years ago – well before modern humans arose. This is consistent with the indications from the fossil record of the break being at least 300,000 years ago, since older Neanderthal-like fossils may still lie undiscovered, and the mitochondrial mutation rate, or "clock" is likely uncertain by a factor of at least two. Indications that the modern individuals of European and Asian stock share more of the Neanderthal genome than do modern Africans indicates that some amount of interbreeding occurred between modern humans and Neanderthal after modern humans had migrated out of Africa. However, analysis of these genes suggest the interbreeding occurred before modern humans entered Europe and the more distant parts of Asia. Once moderns had found

their way well into Asia and Europe, the story of *Homo neanderthalensis* remains a largely separate one from our species, played out on the same stage at the same time.

20.7.3 Neanderthal lifestyle

Neanderthal cultures have been exaggerated in the popular literature in both directions – emphasizing the primitive and exaggerating their abilities. Neanderthals buried their dead, but the extent to which the burials were ceremonial remains in dispute. (The arrangement of artifacts and animal bones is not much removed from accidental, in most cases.) They left no cave paintings, unlike the prolific European artists, Cro-Magnons, who replaced them, but the Neanderthals did leave evidence that they used pigments for some purposes. They had distinctive tool styles, yet variety and innovation are extremely limited: Neanderthal tool types remain similar for blindingly long expanses of time (tens of thousands of years). The sophistication of the tools, compared to those of Cro-Magnon, is low and would have provided relatively limited assistance in a physically demanding environment.

In some cases, a handful of different tool styles will be seen in a limited area (about 100 km in extent) for thousands of years. This, combined with other evidence that Neanderthal population densities were always very low compared with that of modern humans, suggests that Neanderthal populations didn't interact with each other. Groups would come and go across a landscape, rarely or never encountering each other. This is very different from all modern human cultures; modern humans are a traveling species characterized by the continual interaction of different tribes, cultures, and nations.

Part of the reason for such noninteraction may be that Neanderthal groups ranged over very limited areas. Analysis of tools and animal remains suggests that hunting occurred, but not with the reliance on sophisticated tools constructed by even early tribes of modern humans. The extent of skeletal injuries among Neanderthal finds suggest that hunting may have been a very physical and brutal affair: cooperative certainly, but low tech. Foods gathered and scavenged were likely important components of their diet as well.

Details of Neanderthal social life are at best sketchy; at worst, fictional. The anatomy of the skull and neck area suggest that Neanderthals could not be as articulate as modern humans; whether that meant that speech was not heavily employed is unclear. The arrangement of family groups is also speculative. Some anthropologists argue that the characteristics of Neanderthal hearths and other structures in caves imply a very different arrangement from most or all modern humans; in particular, one in which males lived separately from females in day-to-day existence. Other anthropologists argue that such inferences constitute overinterpretation.

At the heart of such musings lies the question of the Neanderthal mind. Given that we do not understand well the nature of our own brain, speculations based on skull size and shape are dangerous ones. Undoubtedly there were differences in the behavior, capabilities, and skills of Neanderthals relative to moderns; unfortunately, the nature of those differences is so faintly hinted at by the physical evidence that they remain wholly mysterious.

20.7.4 Interaction of Neanderthals with moderns

Neanderthals and moderns overlapped in geographic range for almost a third the duration of Neanderthal's existence. Modern forms of *Homo sapiens* moved into the Middle East from Africa by about 90,000 years ago. Neanderthal, under pressure during especially cold periods to move south, is found as early as 120,000 years ago and as recently as about 50,000 years ago in the Middle East.

As modern humans pushed outward from Africa, they began to appear in Europe by about 45,000 years ago, spurred on perhaps by episodes of unusual warmth around that time. Unlike the Middle East, a geographic crossroads from which both species came and went, Europe is a continental cul-de-sac. As moderns spread across Europe, bringing sophisticated tools and weaponry, efficient hunting techniques, and a lifestyle that included much contact and interchange between tribes, the Neanderthals began to be pressured. It would take almost 20 millennia for the Neanderthals to succumb; at any given time it might well have looked like the two species were coexisting peacefully.

A sign of the pressure on Neanderthals is a change in their monotonous stone tool culture. Later tool sets associated with Neanderthals show much more variety than do their earlier classic tool types, and a resemblance to the kind of tool kits the moderns were using. Whether Neanderthal tried to imitate the moderns to gain the latter's hunting advantage, or for other reasons, the change in tool types occurs only after modern-type humans arrived in Europe.

From 40,000 to 27,000 years ago, the geographic range of Neanderthals shrinks progressively, ending in southern Spain. This area is geographically distant from natural migration routes, and represents a logical "last refuge" for a people who are succumbing to whatever pressures the moderns were bringing to bear. Extinction need not have been caused by war or other direct suppression. Only a very small reduction in breeding success is required to eventually drive a species to extinction. For a typical human generational interval (20–30 years), a roughly 2% difference in successful child-rearing between Neanderthals and moderns could have led to Neanderthal extinction in a millennium.

The moderns who first migrated to Europe and, by their advanced hunting techniques and gregarious lifestyles, drove Neanderthal to extinction, were not the Europeans of today. They were Cro-Magnon, a tall and slender race that does not resemble any of the modern peoples of Europe. They were, however, anatomically modern in essentially all respects, and the differences in features from today's Europeans is racial in nature. Successive migrations to Europe over the millennia brought other peoples to Europe; it is possible to trace many such waves just as one can for other parts of the world. The most ancient Europeans living today are thought to be the Basque people, both on linguistic grounds and through analysis of mitochondrial DNA. Well before them, however, came Cro-Magnon and others who have left their legacy in cave paintings, animal sculptures, musical instruments, elaborate burials, advanced tool kits, evidence of highly organized settlements, and perhaps a genetic contribution to later peoples of Europe.

20.7.5 Who were the Neanderthals?

The bulk of the anthropological evidence indicates that Neanderthals were a separate species of humans that evolved more or less in place from earlier *erectus*, or closely related, species. This evolution occurred during a time when various other archaic populations, less well understood from the fossil record, arose from *erectus*-type populations in Africa, Asia, and possibly Australia. The Neanderthal speciation resulted in a people who had a cranial capacity similar to or larger than modern humans, but with significantly different physical and cultural attributes, reflecting perhaps substantial behavioral and intellectual differences as well. Displaced by modern humans who originated much later than they did, the Neanderthals are considered to be a separate and older natural experiment in the speciation of human beings, one that lived a long time and nearly made it to the present day.

The focus here on the Neanderthal story is not meant to imply that it is the most important episode in human evolution. It is, instead, the best documented of the interactions between archaic human populations – those derived from the ancient *Homo erectus* migrations out of Africa – and moderns, those peoples resulting from the much later speciation event in Africa that produced modern *Homo sapiens*. The replacement of archaics by moderns occurred elsewhere around the world (excepting the Americas and Antarctica, where archaics were absent), but nowhere else is the physical evidence so extensive and clear.

We yearn to meet ancestors who will tell us where we came from and why – we people our myths with giants and elves, ogres and trolls, beings who are not quite human, and whose imagined existence allows us to hold a mirror up to ourselves, to evaluate what it truly means to be human. The occasional encounter of modern humans with Neanderthals between 45,000 and 27,000 years ago might have carried with it some of that mythic quality, a reckoning with another intelligent species whose common origin in the distant past could have been intuited by both species but not understood. Those of us alive today missed the chance for such an encounter by no more than a quarter of the span of time of modern humans, and less than 2% of the Pleistocene epoch.

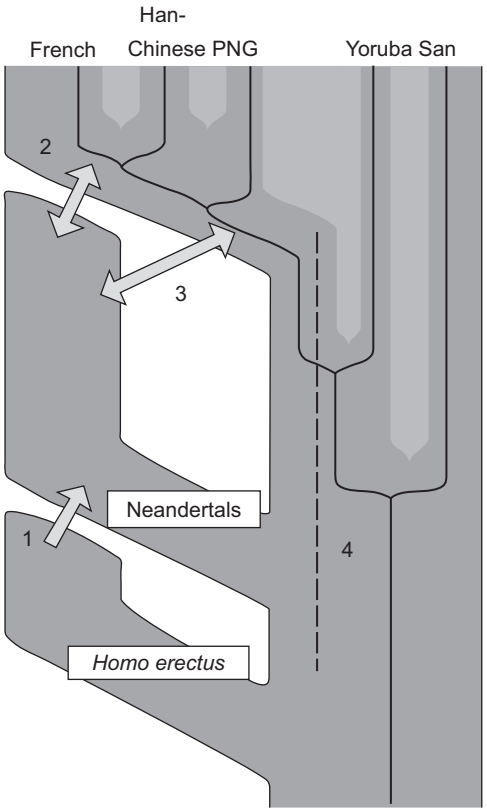


Figure 20.7 Four possible scenarios of genetic mixture involving Neandertals. Scenario 1 represents gene flow into Neandertal from other archaic hominins, here collectively referred to as *Homo erectus.* This would manifest itself as segments of the Neandertal genome with unexpectedly high divergence from present-day humans. Scenario 2 represents gene flow between late Neandertals and early modern humans in Europe and/or western Asia. We see no evidence of this because Neandertals are equally distantly related to all non-Africans. However, such gene flow may have taken place without leaving traces in the present-day gene pool. Scenario 3 represents gene flow between Neandertals and the ancestors of all non-Africans. This is the most parsimonious explanation of our observation. Although we detect gene flow only from Neandertals into modern humans, gene flow in the reverse direction may also have occurred. Scenario 4 represents old substructure in Africa that persisted from the origin of Neandertals until the ancestors of non-Africans left Africa. This scenario is also compatible with the current data.

20.8 This modern world

With the demise of the last archaics – the Neanderthals – modern humans became the singular branch of the hominid family to inherit Earth. Once begun some 100,000 years ago, migrations did not stop, and never have; over and over again from one continent to another waves of human migrants have traveled and made new homes (Figure 20.7). Southeast Asia was reached perhaps 65,000 years ago, from which the first modern humans touched Australian soil 6,000 years before present. Northern Asia did not see modern humans until 25,000 years ago, and the Americas were entered from there no later than 15,000 years before present. The mid-Pacific islands were reached by humans only a few thousand years ago; Antarctica a century ago. The most recent human landfall on a hitherto untouched place was on the Moon in 1969, and we continue to explore new realms of the ocean floor.

The story of Earth takes a new turn with the spread of modern *Homo sapiens*, one in which the progressive growth of agricultural and industrial societies creates novel impacts on land, oceans, and atmosphere. To understand this present time, we must begin in the last glacial episode, with the details of the climate and vegetation record that provide the baseline from which anthropogenic influences can be evaluated.

Summary

Human origins lay in Africa when shifts in climate caused dramatic and repeated changes in landscape types and forest cover. The fossil record shows ape-like animals arising between 5 and 2 million years ago, progressively moving from species not much different from the apes of today to creatures very different from them and yet not human. About two million years ago – as the glacial–interglacial oscillations firmly took hold, the first tool-using members of the genus *Homo* arose in Africa. The most successful and long lived of these, *Homo erectus*, existed for over 1.5 million years and spread beyond the African continent. Based on analysis of the human genome, modern humans arose in Africa between 100,000 and 200,000 years ago, and began their own migration into the Middle East, then Europe and Asia. There they encountered species evolved from more archaic members of the genus *Homo*, including in *Homo neanderthalensis* in the Middle East and Europe, with whom humans would coexist for tens of thousands of years. Neanderthals, despite contributing some genes to modern humans, were a separate species with specialized physiologies for cold weather and with distinct behaviors in terms of hunting and toolmaking. Intelligent but not flexible in their tool kits and culture, their long reign of over 200,000 years across Europe and Western Asia ended less than 30,000 years ago. They were extinguished not by war with moderns but by changes in climate and the effects of competition for resources with the more adaptable moderns. Today, only one subspecies of this long parade of members of the genus *Homo* exists: *Homo sapiens sapiens* – a migratory creature with the intellect and drive to span the globe, and reach into the depths of the oceans and outward to our cosmic neighborhood.

Questions

1. If climate instability stimulated the development of humans, why were not similarly sophisticated creatures the product of earlier epochs of climate instability?
2. Given the fate of the Neanderthals at the hands of humans, what might be humanity's fate should we ever encounter a similar, but technologically more advanced, intelligent species beyond Earth?
3. Although it is almost impossible to isolate humans today for any lengthy period of time, can you imagine a genetic change – even one with social or behavioral consequences – that could procreatively isolate a population of people from the rest of humanity and thus effectively generate a new species?

General reading

Finlayson, C. 2010. *The Humans who went Extinct: Why Neanderthals Died Out and We Survived.* Oxford University Press, New York.

Stringer, C. and Andrews, P. 2012. *The Complete World of Human Evolution*, 2nd edn. Thames and Hudson, London.

References

Can, R. l., Stoneking, M., and Wilson, A. C. 1987. Mitochondrial DNA and human evolution. *Nature* **325**, 31–6.

Kimball, W. H., Johanson, D. C., and Rak, Y. 1994. The first skull and other new discoveries of *Australopithecus afarensis* at Hadar, Ethiopia. *Nature* **368**, 449–51.

Green, R. E., Reich, D., Paabo, S. *et al.* 2010. A draft sequence of the Neandertal genome. *Science* **328**, 710–22.

Larrick, R. and Ciochon, R. L. 1996. The African emergence and early Asian dispersals of the genus *Homo*. *American Scientist* **84**, 538–51.

Macaulay, V., Hill, C., Achilli, A. *et al.* 2005. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science* **308**, 1034–6.

Mellars, P. 1996. *The Neanderthal Legacy*. Princeton University Press, Princeton, NJ.

Stringer, C. and Gamble, C. 1993. *In Search of the Neanderthals: Solving the Puzzle of Human Origins*. Thames and Hudson, London.

Tattersall, I. 1995. *The Last Neanderthal: The Rise, Success, and Mysterious Extinction of Our Closest Human Relatives*. Macmillan, New York.

Tattersall, I., Delson, E., and Van Couvering, J. 1988. *Encyclopedia of Human Evolution and Prehistory*. Garland Publishing, New York.

Thorne, A. G. and Wolpoff, M. H. 1992. The multiregional evolution of humans. *Scientific American* **266**(4), 76–83.

Waddle, D. M. 1994. Matrix correlation tests support a single origin for modern humans. *Nature* **368**, 452–4.

White, T. D., Suwa, G., and Asfaw, B. 1994. *Australopithecus ramidus*, a new species of early hominid from Aramis, Ethiopia. *Nature* **371**, 306–12.

Wilson, A. C. and Cann, R. L. 1992. The recent African genesis of humans. *Scientific American* **266**(4), 68–73.

PART IV

The once and future planet

21

Climate change over the past few hundred thousand years

Introduction

Humankind's present-day dilemma with respect to global warming often is viewed with virtually no temporal perspective at all. The World Meteorological Organization reports that the decade 2001–2010 was the warmest on record, surpassing 1991–2000, which itself was warmer than previous decades. But how does this century compare to other centuries, or this millennium to others? In the third part of this book, we explored extremes of Earth climate far more profound than those experienced in modern times, or even through the short span of human history.

To really put global warming in perspective, however, we need to understand how the climate has varied during the penultimate geologic epoch, the Pleistocene, a time when all of Earth's geologic processes, and the chemistry of the atmosphere, are fully modern in every respect. The time since the last interglacial, through the last ice age to the present interglacial, is recent enough that evidence is available by which very detailed records of climate can be constructed. The most thorough records can be assembled for the past 10,000 years of Earth history, the Holocene. In this chapter, techniques for assembling detailed climate information are summarized, and we compare the climate in this interglacial with that in the last, a kind of "Jekyll and Hyde" story.

21.1 The record in ice cores

As discussed in Chapter 6, the stable heavy isotopes of both hydrogen and oxygen exist in ocean water, and the resulting heavy water tends preferentially to exist in liquid form as opposed to vapor. Thus, water evaporated from equatorial oceans and moved poleward in storm systems is progressively depleted in heavy water, and this effect is more pronounced in colder climates. The ice sheets deposited in polar latitudes over the past 400,000 years therefore contain a record of warmer and colder times through the amount of depletion of deuterium and ^{18}O in the ice. Together, the hydrogen and oxygen isotopes in ice cores allow a record of temperatures to be assembled with quite high fidelity, showing century or even decadal variation, back through four glacial/interglacial cycles. The ice cores also record carbon dioxide content in the atmosphere at the time each layer is laid down, because carbon dioxide is trapped in air bubbles in the ice during deposition each winter. Other gases that may contribute to the trapping of atmospheric heat are found in the bubbles. The cores also contain a record of the amount and kinds of dust that blew across and were deposited annually on the ice sheets. Glacial epochs seem to be not only colder but also drier, on a worldwide basis, so that broader deserts and hence more airborne dust are a signature of those times.

Ice core records dating back through several glacial cycles must be collected from sites that have remained glaciated throughout that time to the present. High latitude or high altitude is required for persistent glaciation. However, many such sites are very dry, and hence the ice layers deposited are thin. Pressure from the continuing addition of annual layers eventually squeezes the layers to the point where they cannot be sampled. Periods of warmth cause a diferent problem: the diffusion of oxygen and other isotopes through the softening ice eliminates the annual layers and may even smooth out the decadal or century-scale variations. Furthermore, correlating core depth with dates is not easy. For cores in which annual ice accumulation is large, the annual cycles may be counted directly. Nearer the bottom of cores or in drier regions, the annual variations are smeared out and a model of ice accumulation that is tied to the inferred temperatures must be applied, or sea core sediments can be used to correlate ages.

Two regions of Earth that have produced excellent records are Antarctica and Greenland; their positions in opposite

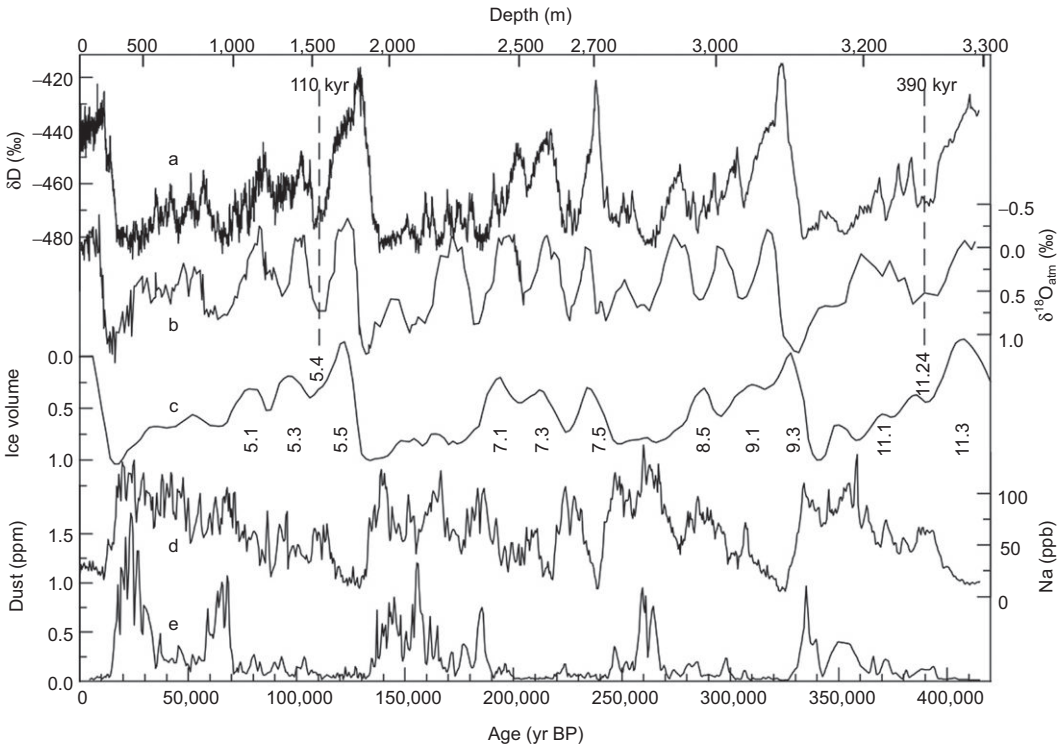


Figure 21.1 A Vostok ice core extending over 3.6 km depth showing various isotopic and other indicators of climate. Amount of deuterium relative to hydrogen in the ice is a measure of ocean temperatures; higher deuterium (lower negative number) means higher temperature at the time of deposition. Oxygen isotopes (18 and 16) are also shown; smaller values (upward) indicate warmer conditions. (δD in per mil means the difference between D/H in the sample, and D/H in present-day ocean water, normalized by the latter and multiplied by 1,000; $\delta^{18}O$ is defined the same way, but with respect to ^{16}O.) Other parameters include ice volume, amount of dust, and sodium (Na) content. This last is a measure of ocean storminess, since the sodium comes from sea salt wafted by ocean spray. Reproduced from Petit *et al*. (1999) by permission of Macmillan Magazines, Limited.

hemispheres of Earth have allowed a determination of how widespread various climate changes might be. Figure 21.1 is an ice core from the Vostok station in Antarctica showing 3.6 km of ice core. Four glacial cycles are represented in the data in the figure corresponding to over 400,000 years. For comparison, an ^{18}O record from seafloor sediments is shown, and the two track each other very well. The ice core, however, clearly is more detailed, showing shorter duration variations. The ice core temperature record also tracks the carbon dioxide record, as shown in the Vostok core in Figure 21.2. Lower carbon dioxide values seem to correspond to lower temperatures. Whether the carbon dioxide is responding to, or forcing, the temperatures is a key puzzle in the study of Pleistocene climates that we return to in Chapter 22. The carbon dioxide record is much less accurate than the isotopic record because of the problems of diffusion of the carbon dioxide through the ice. It is very important, however, not just for correlating temperature changes with carbon dioxide variations, but also because of the possible direct effects on plant communities of changes in the carbon dioxide content of the atmosphere.

The basic pattern over the past 100,000 years or so begins as the last interglacial the Eemian interglacial gives way at 115,000 years before present to the last of the Pleistocene glacials. An initial period of extreme cold, blurred in the sea sediment record, rapidly retreats and a mild glacial time oscillates in warmth, until a second deep glacial some 60,000 years ago is reached. Climate then moderates, but cools again progressively with oscillations seemingly on all timescales resolvable by the core until 19,000 years ago when the glacial climax is reached. The glacial snow line where ice exists year round dropped some 1,000 meters (3,300 feet) from today's value, and glaciers pushed down through much of northern Europe, Asia, and North America. Glaciers were even present in some mountainous parts of North America equatorward of 35° latitude. Some 5,000 years later, temperatures began to rise quickly, and the present interglacial began.

Careful examination of the ice core record in Figure 21.1 reveals that the Eemian and Holocene interglacials are different in their character. The onset of the Eemian 135,000 years ago is characterized by a time of extreme warmth, exceeding anything in the Holocene, and an apparent progressive decline through average Holocene levels until the precipitous drop off into the glacial. The Holocene is characterized by an equally sudden rise, but to a value only somewhat above the average temperature for the past 10,000 years. Following this rise, the temperature seems to settle to a plateau that is broken only occasionally by modest excursions. This interglacial, the one in which human civilization began and has flourished, appears to be more stable than the previous one. The contrast between Holocene and Eocene conditions differs in different ice core data. In some data there is little difference between the two climates, but the Eemian–Holocene difference is striking in the higher resolution ice core

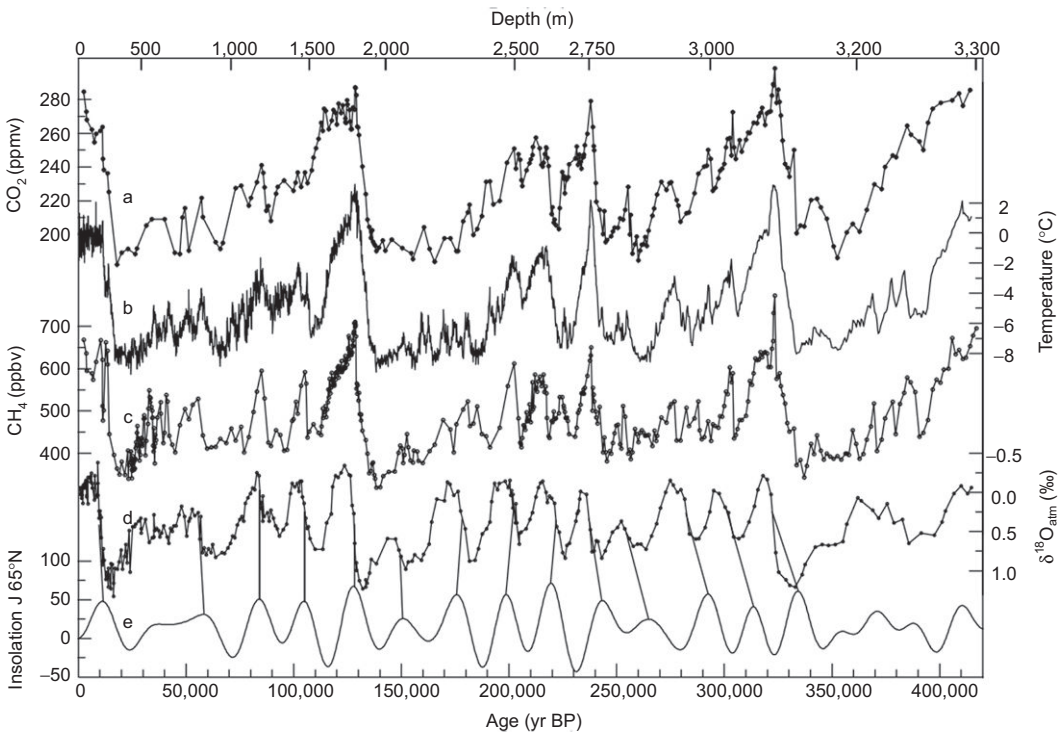


Figure 21.2 A record from the Vostok ice core showing atmospheric carbon dioxide and methane abundances and temperature $\delta^{18}O$ over the past 420,000 years. Temperature is derived from ^{18}O in sediments and deuterium isotopic abundance in the ice. Also shown is $\delta^{18}O$ of oxygen trapped in air bubbles in the ice, a proxy not of temperature but of global ice volume. Finally, the amount of sunlight in June at a particular (northern) latitude is given to show the effect of spin and orbit Milankovitch cycles discussed in Chapter 19; the authors conclude that there is a strong imprint of these cycles on the ice volume data. Reprinted from Petit *et al*. (1999) by permission of Macmillan Magazines, Limited.

record collected in Greenland. In that moister climate, the annual deposition of ice is greater, hence the resolution of the temperature record is higher. However, to reach the Eemian requires drilling deeper. The start of the Eemian is almost 2.9 km below the surface in the Greenland core, by which depth the Vostok core has reached twice that age. However, the former's resolution is remarkable, as shown in oxygen isotope data from the Greenland core (Figure 21.3).

The apparently different character of the Eemian interglacial climate has been challenged on the basis of evidence that the Greenland ice cores showed distortion near their bases; some of the extreme variability might therefore be an artifact. This issue is not fully resolved. European pollen core data (see next section) seem to support the higher variability of the Eemian climate, and indicate warmer summers in Europe than occur throughout the Holocene. The warmer temperatures during the early part of the Eemian are certainly not in dispute. Fossil evidence, such as the occurrence of tropical animals, suggests that northern hemisphere winters during the Eemian were warmer and wetter than now. Other data and modeling suggest significantly higher sea level at the Eemian peak, perhaps by 4 to 6 meters.

21.2 Climate from plant pollen and packrat midden studies

Plants are sensitive indicators of climate. Different plants thrive in different locations. Dry, cold conditions might produce grassy steppes whereas a change to warm and wet circumstances will bring trees to the same area. In Europe and similar climates, cold conditions favor evergreen trees while in warmer periods deciduous trees are more common. In the American Southwest and other arid but mountainous parts of the world, vegetation is a sensitive function of altitude: desert plants give way to scrub woodlands and then to forests as one climbs upward from plains to mountain heights. While separating the effects of rainfall from temperature is not easy, plants have proved to be a unique means of directly sampling local and regional climate changes.

Plant remains are surprisingly durable if they are stored under the right conditions. Sediments on lake bottoms or in dried lake beds preserve plant pollen in conditions that are often anoxic, or otherwise favorable for the long-term preservation of material. Dating of the material through the Holocene is possible using the radiocarbon technique, i.e., determining the $^{14}C/^{12}C$ ratio to infer a date, as described in Chapter 5. For earlier times the dates become increasingly uncertain, and for the Eemian (much more ancient than the 70,000-year useful limit for such techniques) less precise estimates must suffice.

Figure 21.4 is an example of the occurrence of pine and alder pollens in lake bottom sediments in France and Germany, with ages dating to the Eemian. Fluctuations in occurrence of pollen, if they translate with fair fidelity to climate, suggest that the Eemian was indeed characterized by strongly variable climates. During warmest Eemian times, temperatures may have been 5°C higher in the winter than on average today in Europe, and the coldest of Eemian episodes were comparable to those in glacial

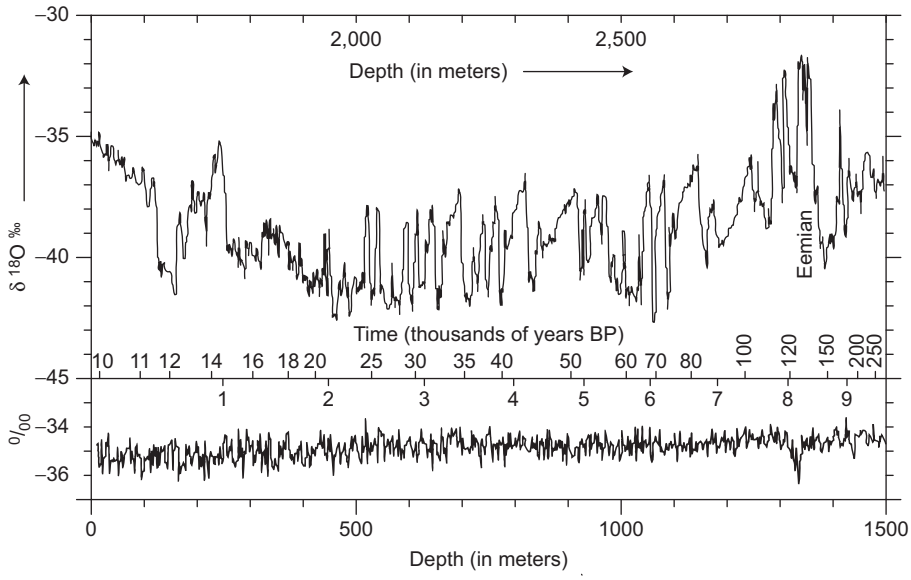


Figure 21.3 Greenland ice core showing ^{18}O concentrations versus time. Colder periods are characterized by less ^{18}O (more negative numbers) because relatively more ^{16}O is being sequestered from the oceans into the ice. Time in the past, corresponding to depth in the core, is labeled on the horizontal axis within the panel. For compactness, the core is broken up into 1.5-km segments: the lower figure covers the past 10,000 years; the upper figure represents core material much more compressed by the weight of the ice overburden, hence extending back 250,000 years. Adapted from Dansgaard *et al*. (1993) by permission of Macmillan Magazines, Limited.

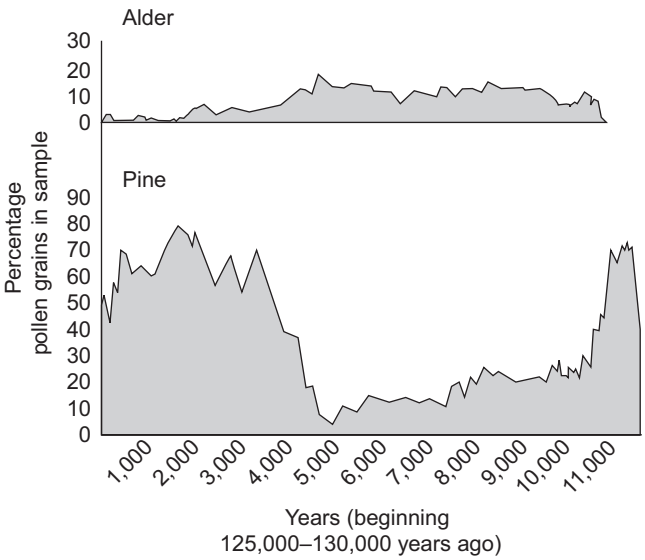


Figure 21.4 Occurrence of pine and alder pollens in lake-bed sediments from the Eemian interglacial in France and Germany. More pines suggest colder conditions; more alders, warmer conditions. Adapted from Field *et al*. (1994) by permission of Macmillan Magazines, Limited.

conditions. Further, the extent and timing of variations may have differed in different parts of Europe, and hence almost certainly differed in geographically more distant locations.

Pollen studies also have been used to track climate changes in continental locations as the last ice age came to a close. In arid regions of the world, such as the southwestern United States and northern Mexico, these studies are supplemented by information from packrat middens.

Packrat middens are well-preserved fragments of plants, and less commonly insects, accumulated locally by packrats (also called woodrats). Their preservation is the result of being encased, amber-like, in crystallized urine, a reflection of the habit of packrats to urinate on their caches of collected material. The dry climate of the Southwest can preserve this *amberat* for tens of thousands of years. Dating is done by radiocarbon techniques on the organic material. Packrat middens are common enough in caves and crevices throughout the southwestern United States that a useful climate record over 40,000 years has been assembled through studies beginning around 1960. The same kind of studies have more recently been performed in South America, Africa, the Middle East, Asia, and Australia.

The source of the climate information lies in the fact that packrats collect material from a relatively limited area around a midden site, and enough is collected over the life of the animal that the midden becomes representative of the local plant environment at the time the animal lived. Corrections must be made for the variation in types of plants collected by different packrat species (a function of their diet).

The result of dated amberat samples is a map in space and time of the plant communities, and hence climate shifts, of a given area, dating back halfway through the last ice age. Figure 21.5 shows an example of this kind of work in the form of vegetation maps of the southwestern United States. This region, stretching from California to western Texas, encompasses large parts of three great Mexican and American deserts: Mohave (California), Sonoran (Arizona), and Chihuahuan (New Mexico and Texas). They differ in their elevation and amount and seasonal timing of rainfall. The deserts give way in the uplands and mountains

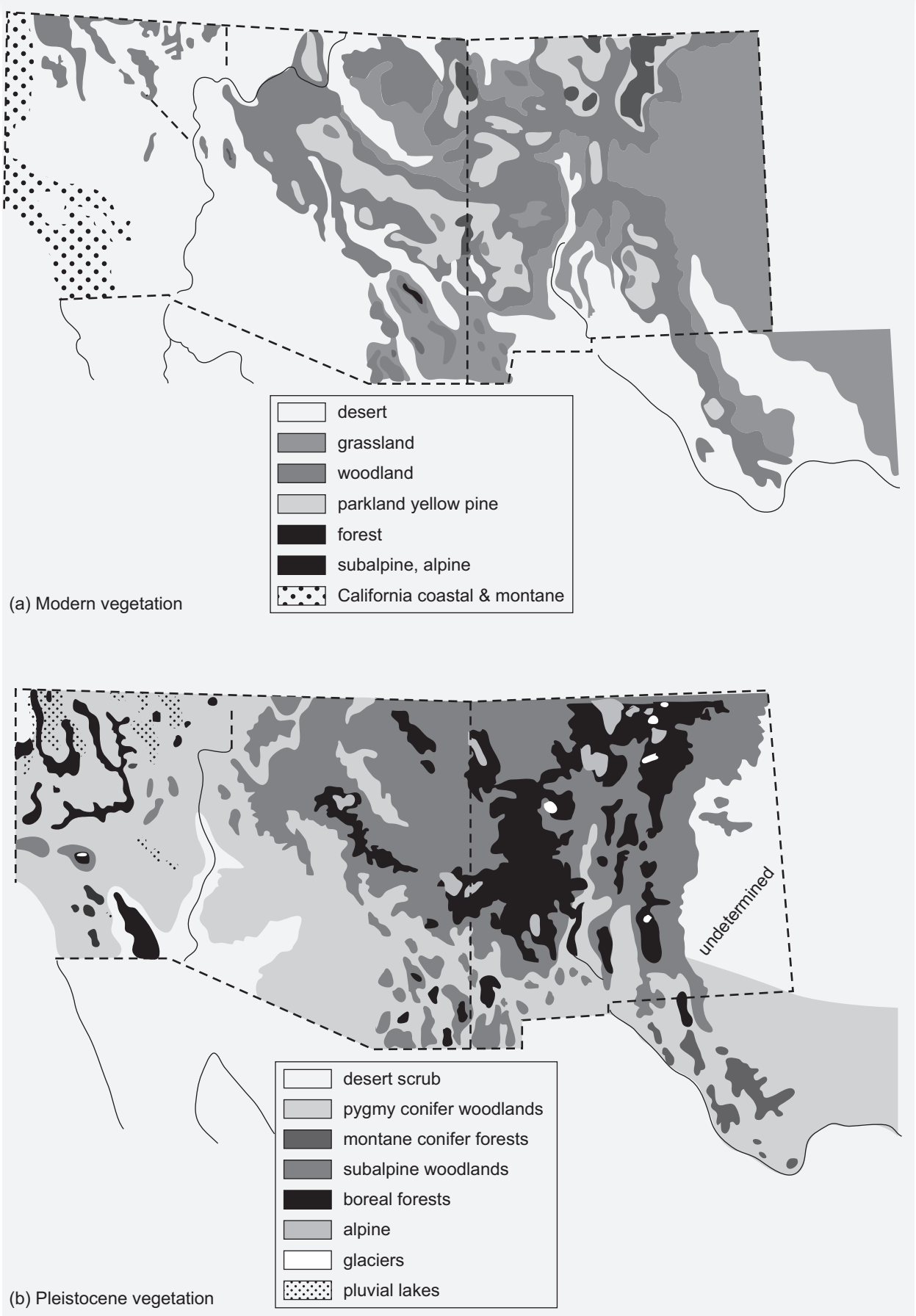
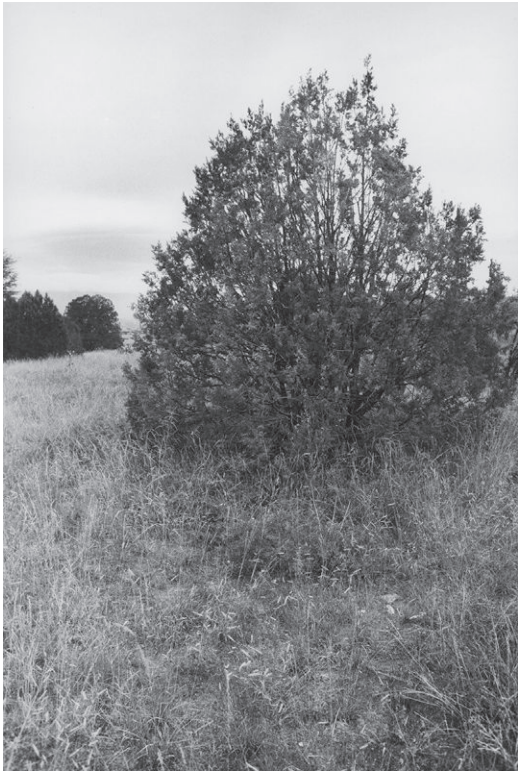


Figure 21.5 (a) Map of southwestern United States vegetation from today and (b) a similar map from 18,000 years ago, based on packrat midden and pollen studies. Note that keys are slightly different, in part because vegetative associations and species types changed in addition to the altitude of occurrence. Roughly, "parkland" and "forest" in (a) correspond to "montane conifer forests" in (b). Panels (a) and (b) are based on Betancourt *et al*. (1990). (c) Typical desert vegetation seen today in the Sonoran desert of Arizona, such as the area around Tucson. Panel (c) is photograph by the author. (d) Juniper trees some 80 km southeast of, and 600 meters above, Tucson. Junipers similar to these (though of a different species) were present in the Tucson basin during the last ice age. Panel (d) is photograph by the author.

(c)



(d)

Figure 21.5 (*cont.*)

to other plant communities as a function of altitude and rainfall, which can differ greatly depending on the orientation of the mountain ranges. Roughly, though, today's grasslands appear above 1,000 meters, woodlands above 1,500 meters, pine parklands above 2,000 meters, conifer forests above 2,500 meters, spruce-fir forests above 3,000 meters (subalpine conditions), and tundra above treeline (alpine) above 3,500 meters.

In the depths of the ice ages, conditions were different. The same region 18,000 years ago looked as if one had raised the general altitude of the land 500 meters or so. Tucson, Arizona, at an elevation today of 700 meters, is in the Sonoran desert; 18,000 years ago, Juniper and Pinyon pine woodlands interspersed with grasslands occupied the Tucson basin. The Santa Catalina Mountains north of Tucson, where today big conifers grow typically above 2,000 meters, contain ice age packrat middens with conifer seeds at 1,500-meter altitude. The upper mountain reaches of the Catalinas, just shy of 3,000 meters, were "too high" then for trees; today that summit is covered with Douglas fir and White fir trees.

An important complexity in this story, revealed by the amberat data, is that the change from glacial to interglacial was not simply a matter of moving the same vegetative communities up in altitude; latitudinal shifts occurred as well. Species that were not adapted to glacial climates became more abundant in the Holocene, and some species widespread in the last glacial were squeezed out by new plant species. Many plant types of subtropical origin only colonized the Sonoran desert of southern Arizona and northern Mexico during the Holocene. In the nearby mountains, Douglas fir did not have to compete during the last glacial with Ponderosa pine, a tree that today occupies much the same elevation regime. Ponderosas apparently are particularly adapted to the moisture pattern of the Holocene, and moved southward to colonize large areas of the Arizona uplands only after the last glacial ended.

The importance of the packrat midden and pollen records goes beyond the estimation of climate variation based on vegetation occurrence. They show, by example, the tremendous ecological displacements associated with climate changes. Not just a theoretical inference, the massive displacement of vegetative types and associated animals is preserved as physical evidence in lake sediments and packrat middens. Such records are a reminder that, if indeed human activities today are precipitating or accelerating global warming, we should expect significant changes in occurrence of forests, woodlands, and agricultural belts as a result of such activities.

21.3 Tree rings

Many species of trees grow in such a way that in cross-section their trunks display rings, reflecting an annual or seasonal cycle to their growth. In very wet climates, or in soils close to the water table in arid regions, the growth of the trees is little affected by year-to-year variations in precipitation. But in soils on dry hillsides in arid regions, the amount of precipitation in a given year may have a substantial effect on the amount of increase of trunk thickness. In those regions, the thickness of the annual

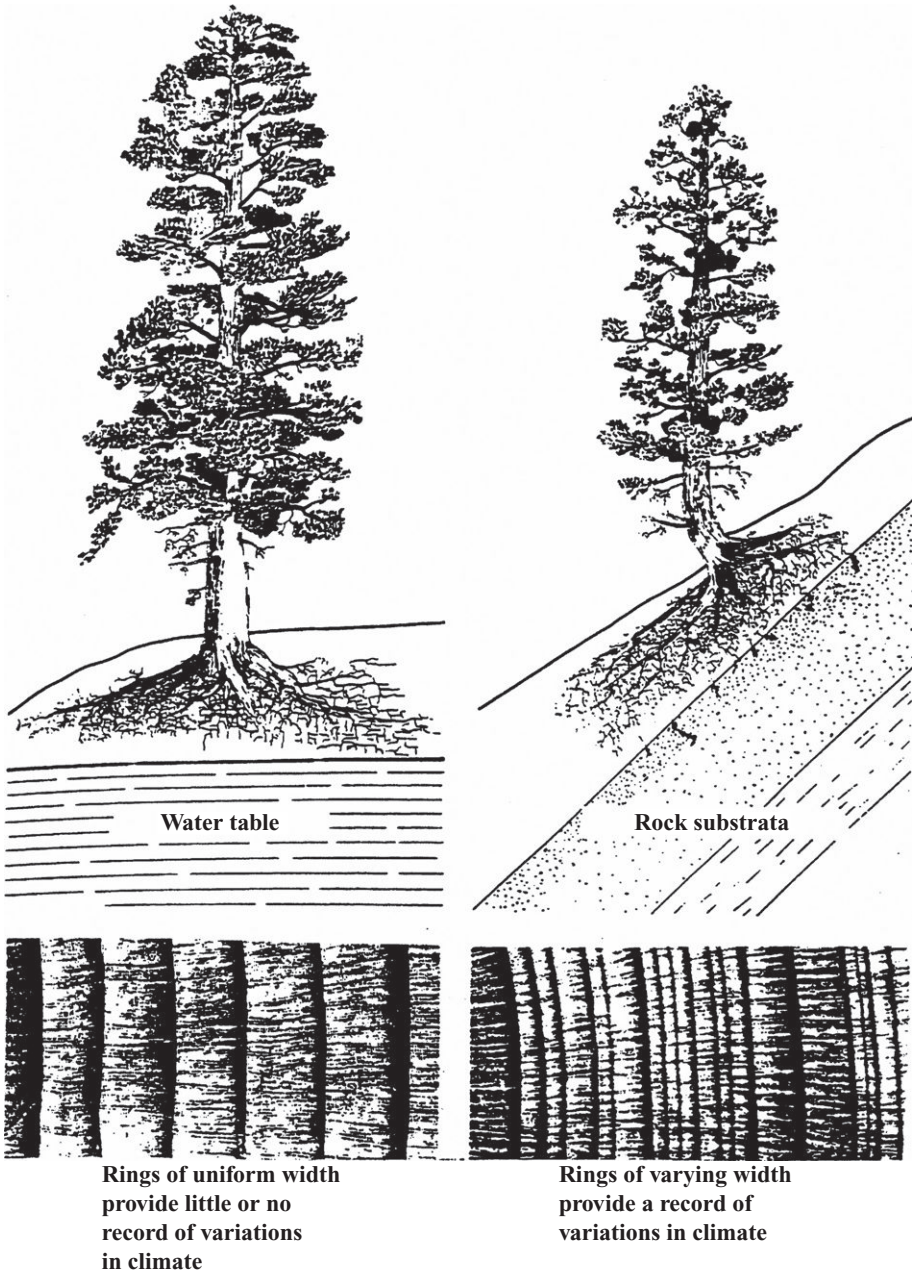


Figure 21.6 Illustration of the appearance of conifer tree rings. On the left, a tree growing where the water table is near-surface is insensitive to annual precipitation variations and will leave regular growth rings. On the right, a tree growing on a dry hillside, dependent on precipitation, will show variations in the thickness of its growth rings, reflecting year-to-year variations in moisture. From the University of Arizona Laboratory of Tree-Ring Research.

or semi-annual growth ring may reflect the precipitation conditions in that locale, that is whether the year was wet or dry (Figure 21.6). Temperature also plays a role in ring thickness, primarily through its effect on soil moisture. Different parts of each tree ring can be examined to provide information on seasonal variations of rainfall within a given year.

Tree rings provide a remarkable record of climate because they show details on timescales as short as seasons. The great age of trees allows an individual tree to record the local climate changes, from one growing season to another, over centuries, even millennia in the case of the Bristlecone pine. It is possible to correlate rings from living to dead trees by matching patterns of narrow and wide rings, or *cross-dating*. By accumulating tree-ring records from a large number of trees in a given region that have experienced a common climate history, one can use overlapping tree-ring patterns from living to increasingly older specimens of dead trees to push the chronology far back in time. By using techniques of statistical analysis on many samples, sources of error such as trees with missing rings or doubled rings (an occasional occurrence) can be minimized. The American astronomer Andrew Ellicott Douglass (1867–1962), who began his tree-ring work at the turn of the century and went on to

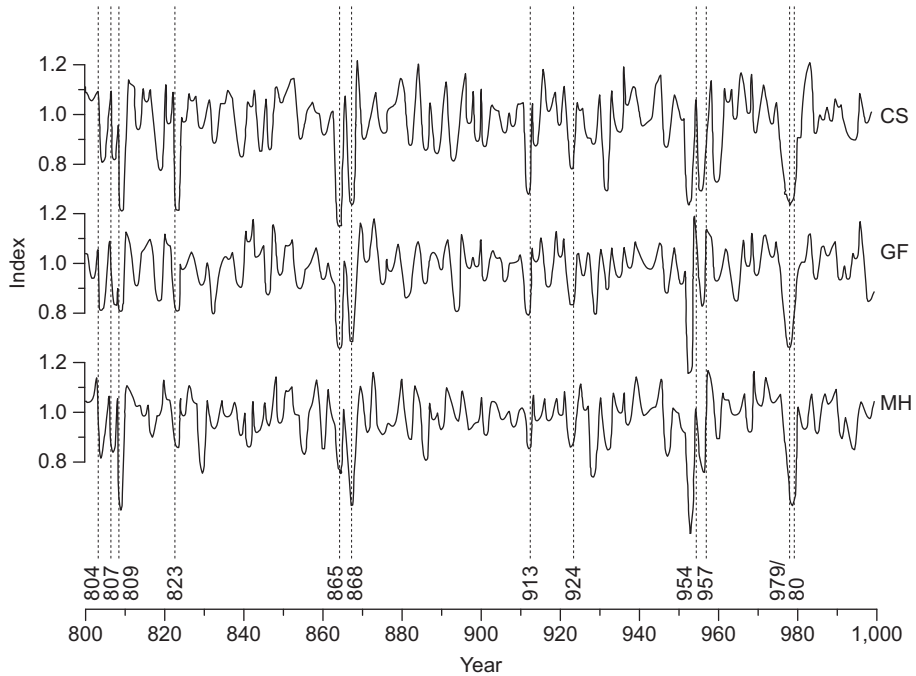


Figure 21.7 Plot of years of wetness and drought in the Sierra Nevada, from A.D. 800 to 1,000. Cores from living and dead sequoias were collected at three sites [Camp Six (CS), Giant Forest (GF), and Mountain Home (MH)] within 100 km of each other. The ring index is a measure of thickness; the higher the index, the thicker the ring and hence the wetter the year. Years in which all three stations show thin rings are marked and are likely to have been ones of severe drought. Reproduced from Hughes and Brown (1992) by permission of Springer-Verlag.

found the pioneering Laboratory of Tree-Ring Research at the University of Arizona in Tucson, first put this technique on a firm quantitative footing.

Because of the great aridity of the mountains in the southwestern United States, samples of Bristlecone pines of enormous antiquity are available, and the tree-ring record of annual climate has been extended back using cross-dating techniques some 9,000 years, 80% of the duration of the Holocene. Some fragments of trees dated by carbon-14 techniques to 11,000 years allow essentially the entire Holocene to be characterized by this technique. Similar efforts are being made at other sites around the world where conifers are available in mountainous climates, for example Tasmania, Chile, Argentina, China, and Tibet. Techniques for coring into living trees without killing them make it possible to sample trees in areas where logging is not permitted.

The tree-ring thickness often is not a simple function of the total annual precipitation in the immediate area, but is a complicated function of several factors. High temperatures in the growing season may increase evaporation and hence soil dryness, even if the year exhibits normal amounts of precipitation. The distribution of precipitation seasonally may be important in affecting the growth of a tree, because the growing season does not extend through the entire year. These lead to ambiguities as to whether a thick ring means a wet year, a cold year but normal in precipitation, etc. A rough, experimental rule of thumb is that high-latitude, high-altitude trees tend to be most responsive to temperature, whereas low-altitude, low-latitude trees (but outside of the tropics) are most sensitive to precipitation.

An example of tree-ring data is shown in Figure 21.7. This data set and others from Giant sequioa trees in the Sierra Nevada of California have been used to create a detailed chronology of droughts in California extending from the present back to 100 BC, a span of over 2,000 years. Time periods when droughts were rare (such as around AD 1,000) are preceded and followed by episodes of repeated droughts, which undoubtedly had a severe effect on the people living in the region at the time. Similar tree-ring data from around the world combine to provide a detailed view of regional climate changes unobtainable with other techniques.

Further information contained in tree rings includes stream-flow data and a history of forest fires, because conifers often survive such fires but retain evidence of the charring that can be dated to a single year. Stream-flow data are obtained by examining annual ring widths in trees growing in the major watershed areas of a given river: wide rings mean wet years with high stream flow, thin ones dry years with low flow. Furthermore, the ability to derive absolute dates from the cross-dating of living and dead trees makes tree rings an important tool for calibrating radiocarbon dates: the age of the wood itself, from the ^{14}C technique, can be compared with the absolute chronology from cross-dating rings.

21.4 Climate variability in the late Holocene

The techniques discussed above are only some of many tools that paleoclimatologists have used to assemble a record of the Holocene climate worldwide. The detail both in time and space

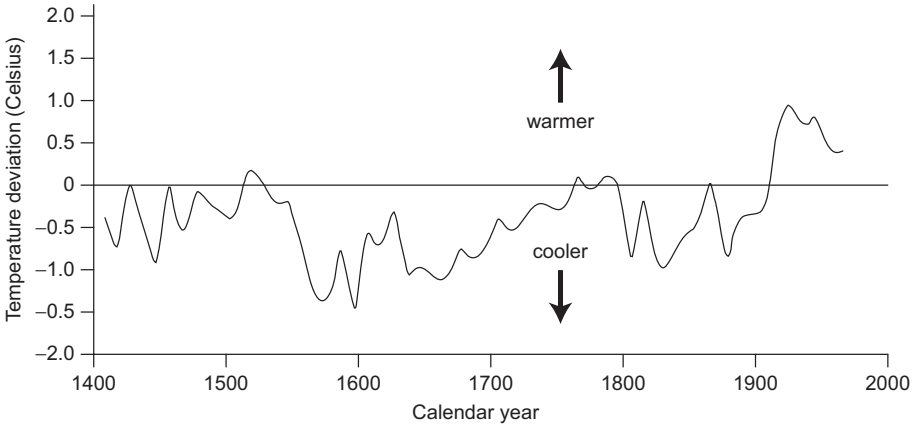


Figure 21.8 Changes in summer temperatures in the northern hemisphere, averaged decade by decade, from medieval times to the present. Deviations from a reference temperature are plotted versus calendar year. The cold period known as the Little Ice Age is evident from the 1600s to the end of the nineteenth century. The figure is drawn from data of Bradley and Jones (1993), who used a variety of information to reconstruct temperatures.

of such a record is unique to this most recent time in Earth's history. Among the recent changes in Earth's climate that have been documented through this record are gradually increasing temperatures following the last deglaciation, the early Holocene Thermal Maximum between 9,000 and 5,000 years ago, a brief cold event of 8,200 years ago, and the African Humid Period, which ended 5,500 years ago with the subsequent aridification of the Sahara desert.

The last major climate perturbation of the Holocene, the Little Ice Age, best documented in Europe, extended from the sixteenth to the early nineteenth century (Figure 21.8). At its peak, annual temperatures dropped 1.5°C from present norms in Europe; in at least some parts of the world it is one of the coldest episodes of the Holocene. During the coldest of those times in Europe there were years with essentially no growing seasons: crops failed, starvation occurred, and rivers known today to remain largely ice-free year round (for example, the Thames through London) often froze over.

A portion of this cold period, from 1645 to 1715, coincides with a well documented and unusual absence of sunspots on the Sun, the so-called "Maunder Minimum." The significance of the sunspot absence for the total brightness of the Sun is unknown, but another indicator of decreased solar activity at the time is a peak in the amount of radiogenic ^{14}C seen in contemporary tree rings. The production of atmospheric ^{14}C depends on the supply of high-energy galactic cosmic rays (Chapter 5). High solar activity pumps up the Sun's magnetic field, which tends to deflect such particles from entering the solar system; low solar activity does the opposite. The increased ^{14}C seen in tree rings from the Little Ice Age is consistent with lower solar activity. This, in turn, suggests that the Sun might have shined less brightly during that time, for reasons that are unknown. A decreased solar output might have been just one contributor to this very recent and unusual cold spell; others include spurts of volcanic activity, or changes in ocean circulation.

Prior to the Little Ice Age, during a time in Europe known as the Medieval Warm Period, major upheavals in agricultural communities were occurring in what is today the southwestern United States. Prolonged periods over a century or two in which the crucial summer rains failed, punctuated by floods that may have damaged or destroyed irrigation canals, seems to be roughly coincident with the decline of organized civilizations that later came to be known as the Anasazi, Mogollon, Hohokam, and others. Whether climate was the direct cause of the dispersal of the peoples who created these cultures, or was only an additional stress on top of others, may never be fully understood.

21.5 The Younger Dryas: a signpost for the oceanic role in climate

Perhaps the most striking episode of climate variability occurred right at the beginning of the Holocene. Named after an arctic flower, the *Younger Dryas* can be seen in pollen-core and ice-core records as a sudden drop in temperature back toward ice-age values, and the resurgence of glaciers, perhaps in less than a century. This cold snap seems to have been restricted to continental areas around the North Atlantic, such as Europe and Canada, and was less significant elsewhere. It lasted a thousand years, and then, in the space of perhaps a decade or two, temperatures rose sharply toward typical Holocene values and the glaciers resumed their retreat.

What is striking about the Younger Dryas is that it occurred during a time when the glaciers were still in the process of retreating from lands that in the full flush of the Holocene are entirely unglaciated. Although the shrinking of the glaciers themselves, in reducing continental albedos, should have reinforced the warming, there are other effects of the retreat that could have played more unpredictable roles.

Geoscientists W. S. Broecker of Columbia University and G. H. Denton of the University of Maine have identified the Younger Dryas as a kind of "smoking gun" for the importance of oceanic–atmospheric interactions in determining climate. During the retreat of the glaciers that began 14,000 years ago, glacial meltwater in North America was accumulating in

southern Manitoba and flowing down the Mississippi River in an impressive torrent comparable to today's massive Amazon River. To the east, glaciers blocked freshwater flow into the North Atlantic. About 11,000 years ago, the ice to the east had retreated sufficiently that much of the meltwater began flowing eastward across what would become the Great Lakes, along the St. Lawrence River and into the Altantic Ocean.

This massive influx of freshwater diluted the normally salty water in the upper layers of the North Atlantic. Water with dissolved salt is denser than fresh water; this is why we float so much better in the ocean than in a swimming pool. In the Atlantic Ocean today, a current of water flows northward at depths of about 800 meters below the ocean surface, until it reaches roughly the latitude of Iceland. There surface winds blow colder surface water aside, the warm water at 800 meters rises, releasing heat. Importantly, it then *sinks* because of its relatively high salinity, creating a "thermohaline circulation" – and, effectively, a heat pump – whereby heat is delivered to the North Atlantic Ocean and the water that delivers it then sinks to great depths.

Because the moderate climate of Europe is dependent on the release of heat from the North Atlantic waters, this Younger Dryas influx of freshwater may have slowed or stalled the thermohaline circulation thereby having a profound chilling effect on the climate for as long as the substantial flow of freshwater continued to pour into the North Atlantic. An alternative model posits a shift in the jet stream associated with the melting ice sheets, delivering more rain to the North Atlantic – with essentially the same effect on the North Atlantic circulation.

The oceanic–atmospheric connection in climate should not come as a surprise. The mass of Earth's oceans is some 500 times that of the atmosphere; the oceans are therefore capable of holding vastly larger quantities of heat than the atmosphere. Similarly, the oceans can hold 60 times as much carbon dioxide as is in the atmosphere today. What has prevented climatologists from understanding the role of the oceans in climate is the lack of knowledge of ocean circulation, and the inherent difference in circulation timescales between ocean and atmosphere. Much of the deep ocean may not mix with the shallower waters on timescales of interest to human global warming issues (decades): just how much does and how it does so are critical to understanding the interaction between the atmosphere and the ocean.

Much of the deepest insight into how the ocean exchanges heat, carbon dioxide, and other important climate quantities with the atmosphere has come from trying to understand the puzzle of glacial cycles: why does Earth become glaciated, why do interglacials occur, and what is the role of carbon dioxide?

21.6 Into the present

The end of the last ice age came as summer sunshine in the northern hemisphere began to approach a maximum (8% greater at 11,000 years ago than at present), the result of the orbital swings of the Milankovitch cycle described in Chapter 19. This may have been the stimulus for a series of changes in ocean circulation patterns that led to worldwide, contemporaneous retreat of the glaciers. The precise mechanism by which the increased northern hemisphere summer heating triggered oceanic changes remains unknown, but the release of additional carbon dioxide stored in the seas represents part of the answer. Worldwide warming of the ocean and melting of glaciers during the early Holocene caused sea level to rise by roughly 130 meters relative to its value at the peak of the glaciation. Higher sea levels reduced the amount of continental shelf exposed above the sea, isolating continents previously connected by land bridges, and contributed to changes in regional climate patterns and migration routes.

As climate changed around the world, the vegetation and animal life changed with it. Large animals began to disappear from widespread regions of the continents, existing extensively now only in Africa and parts of Asia. The relatively stable Holocene climate allowed elaborate forms of agriculture to be invented by humans on all continents. Cities grew up as agricultural and trade centers, perhaps first in the Middle East around 8,000 years ago, then in Europe and the Americas some 2,000 to 3,000 years later. Human population numbers increased steadily as new agricultural techniques and improved transportation technologies were invented. In the past few centuries, humans have harnessed reserves of hydrocarbons trapped in the sedimentary layers of Earth as sources of energy. By the early twentieth century, the use of such fossil fuels was prodigious and had a measurable effect on the total carbon dioxide in the atmosphere. By the early twenty-first century effects of atmospheric carbon dioxide increase on the climate became a matter of deep worldwide concern. An examination of the scientific and human issues behind this debate is the focus of the next chapter.

Summary

Ice cores contain a detailed record of the shifts in climate over the past several hundred thousand years through variations in isotopic ratios of deuterium and oxygen, and can also be used to create a profile of carbon dioxide through time, storminess, and levels of dust in the atmosphere. Only a few places in the world – notably Greenland and Antarctica – have ice sheets that are sufficiently thick and persistent to provide such a climate record through the warmest periods of the last few hundred

thousand years. Some of the striking aspects of the record are the large variability in climate during the glacials, and the relatively stable interglacial climates – particularly the most recent one, the Holocene. The preceding interglacial, the Eemian, started out much warmer than the average Holocene climate. Additional sources of information on climate on hundred-thousand year timescales are isotopes in seafloor sediments, and pollen in lake sediments. On timescales of tens of thousands of years, packrat middens well preserved in arid climates provide a proxy record of climate through the plant types collected by packrats at any given time. On timescales of thousands of years, the record in tree rings allows for an almost year-by-year assessment of variations in climate through the ability to overlap fragments of trees that are now dead with those still living – the Bristlecone pines provide such a record through most of the Holocene. Two striking climate events of the current interglacial are the Younger Dryas, a time when the warming climate suddenly reversed and became cold, and the much more recent Little Ice Age. The Younger Dryas may have been triggered by changes in Atlantic Ocean salinity as North American glaciers melted. The cause of the Little Ice Age, a distinct cooling over several episodes beginning mid-sixteenth century and ending in the nineteenth century, remains uncertain.

Questions

1. How might one use tree rings in a forest of different species of conifers to infer the outbreak of a large insect infestation sometime in the past?
2. What flora and fauna existed in your home area during the coldest part of the Pleistocene?
3. A recent alternative model for the cause of the Younger Dryas invokes the impact of a small cometary or asteroidal fragment. Do a literature search on the web to find information on this model, and discuss its pros and cons.
4. Modern humans arose in Africa, according to the evidence presented in Chapter 20, sometime prior to 100,000 years ago. Thus, most of our lifespan as a species has been on a glaciated Earth. Compare the climate record to the human migration record, and discuss whether there is a correlation. Also, what is it about the behavior of clime during the glacials that might have discouraged agriculture, even in ice-free areas?

General reading

Ward, P. D. 1996. *The Call of the Distant Mammoths*. Copernicus (Springer-Verlag), New York.

References

Berger, A. and Loutre, M. F. 1991. Insolation values for the climate of the last 10 million years. *Quaternary Science Reviews* **10**, 297–317.

Betancourt, J., Van Devender, T. R., and Martin, P.S. 1990. *Packrat Middens: The Last 30,000 Years of Biotic Change*. University of Arizona Press, Tucson.

Bradley, R. S. and Jones, P. D. 1993. Little Ice Age summer temperature variations: Their nature and relevance to recent global warming trends. *The Holocene* **3**, 367–76.

Broecker, W. S. 2006. Was the Younger Dryas triggered by a flood? *Science* **312**, 1146–8.

Brown, P. M., Hughes, M. K., Baisan, C. H., Swetnam, T. W., and Caprio, A. C. 1992. Giant sequoia ring-width chronologies from the central Sierra Nevada, California. *Tree-Ring Bulletin* **52**, 1–14.

Coleman, S. 2007. Conventional wisdom and climate history. *Proceedings of the National Academy of Sciences of the USA* **104**, 6500–1.

Crowley, T. J. 2000. Causes of climate change over the past 1000 years. *Science* **289**, 270–7.

Crown, P. L. 1990. The Hohokam of the American Southwest. *Journal of World Prehistory* **4**, 157–256.

Dansgaard, W., Johnsen, S. J., Clausen, H. B. *et al*. 1993. Evidence for general instability of past climate from a 250-kyr ice-core record. *Nature* **364**, 218–20.

Field, M. H., Huntley, B., and Muller, H. 1994. Eemian climate fluctuations observed in a European pollen record. *Nature* **371**, 779–83.

Greenland Ice Core Project Members. 1993. Climate instability during the last interglacial period recorded in the GRIP ice core. *Nature* **364**, 203–7.

Hughes, M. K. and Brown, P. M. 1992. Drought frequency in central California since 101 B.C. recorded in giant sequoia tree rings. *Climate Dynamics* **6**, 161–7.

Hughes, M. K., Touchan, R., and Brown, P. M. 1996. A multimillenial network of giant sequoia chronologies for dendrochronology. In *Tree Rings, Environment and Humanity* (J. S. Dean, D. M. Meko, and T. W. Swetnam, eds), Radiocarbon, University of Arizona, Tucson, pp. 225–34.

Kaspar, F., Norbert, K., Cubasch, U. and Litt, T. 2005. A model-data comparison of European temperatures in the Eemian interglacial. *Geophysical Research Letters* **32** CiteID L11703.

Loaiciga, H. A., Haston, L., and Michaelsen, J. 1993. Dendrohydrology and long-term hydrological phenomena. *Reviews of Geophysics* **31**, 151–71.

Maslin, M. 1996. Sultry interglacial gets sudden chill. *EOS* **77**, 353–4.

Mayewski, P. A., Maasch, K., Yan, Y. *et al*. 2004. Holocene climate variability. *Quarternary Research* **62**, 243–55.

McCulloch, M., Mortimer, G., Esat, E. *et al*. 1996. High resolution windows into early Holocene climate: Sr/Ca coral records from the Huon Peninsula. *Earth and Planetary Science Letters* **138**, 169–78.

Overpeck, J. T., Otto-Bliesner, B. L., Miller, G. H., Muhs, D. R., Alley, R. B., and Kiehl, J. T. 2006. Paleoclimatic evidence for future ice-sheet instability and rapid sea-level rise. *Science* **311**, 1747–50.

Petit J. R., Jouzel J., Raynaud D. *et al*. 1999. Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* **399**, 429–36.

Steadman, D. W. and Mead, J. I. (eds). 1995. *Late Quaternary Environments and Deep History: A Tribute to Paul S. Martin*. The Mammoth Site of Hot Springs, South Dakota: Scientific Paper No. 3, Hot Springs, SD.

Swetnam, T. W. 1993. Fire history and climate change in giant sequoia groves. *Science* **262**, 885–9.

Tudge, C. 1996. *The Time Before History*. Touchstone Books, New York.

Vostok Project Members. 1995. International effort helps decipher mysteries of paleoclimate from Antarctic ice cores. *EOS* **76**, 169.

World Meteorological Organization. 2011. *The status of the global climate in 2010*. WMO no. 1074, Geneva, Switzerland.

22

Human-induced global warming

Introduction

One of the most fiercely debated social issues grounded in science today is whether humans are affecting the climate of the planet on which we live. While the basic question is a scientific one, the implications potentially touch every aspect of our lives. Are we facing, for the first time in human history, a planet-wide transformation of our environment wrought by human activities? In this chapter the evidence and mechanisms are discussed along with the potential impacts of humankind's effect on climate.

22.1 The records of CO_2 abundance and global temperatures in modern times

Ice cores contain trapped bubbles of air, which, provided they can be properly dated, represent a record of the composition of air over time. Because of the weight of overlying layers of ice, compressing the pores in the ice, it is very difficult to extend the record back as far as that for temperature derived from the isotopic composition of the water itself. In fact, the manner in which the air bubbles were originally trapped in ice results in their movement upward or downward relative to the ice itself, making age determination a challenge.

Figure 22.1 displays CO_2 values from an ice core collected in Greenland. The dating of the air was achieved by taking advantage of a byproduct of nuclear weapons testing: the isotope ^{14}C reached a peak in Earth's atmosphere, from the detonation of nuclear bombs, in 1963. Using this peak in heavy carbon, geochemists M. Wahlen of Scripps Institution of Oceanography and colleagues determined that the trapped air was displaced by the equivalent of 200 years relative to the ice surrounding it.

With this important correction, the figure shows that, during the Little Ice Age, CO_2 values were fairly constant. Beginning in the mid-1800s, carbon dioxide began to increase. Direct measurements from a station in Hawaii, selected to be high above any local industries and hence sampling worldwide CO_2 borne by the trade winds, show that the increase accelerates after World War II.

Today, the carbon dioxide abundance is nearly 40% higher than it was during the Little Ice Age. Some of the increase, particularly that in the mid-nineteenth century, may be ascribed to the general warming that occurred as climate moved out of the Little Ice Age; other ice cores suggest that CO_2 20,000 years ago (near the last major ice age peak) was half that at present. However, some of the nineteenth century CO_2 increase also was likely caused by changes in land-use patterns, including deforestation: during their lifetime, trees are an important *sink*, or removal agent, of atmospheric carbon dioxide.

In the twentieth century, there is little disagreement that industrial activities, that involve the burning of carbon-rich *fossil fuels* (see Chapter 23), are primarily responsible for the increased atmospheric carbon dioxide. Here, industrial is defined broadly to include use of automobiles, home heating systems, as well as agriculture involving burning of forests for clearing. Adding all of these activities together, one expects an even larger atmospheric carbon dioxide increase than is seen in the top panel of Figure 22.1; some of the excess likely is taken up in the oceans and perhaps forested regions of the continents.

Other atmospheric greenhouse gases have increased during the twentieth century relative to preindustrial values. Methane (CH_4) is 2.5 times the preindustrial value. Again, this increase seems most readily accounted for by increased industrial activity and development of intensive agricultural techniques. Nitrous oxide (N_2O) is 20% higher than in preindustrial times; chlorofluorocarbons (CFCs) used in air conditioning and other applications have no natural sources and are appearing in the atmosphere for perhaps the first time in Earth's history. These and other compounds represent a significant perturbation to the background composition of Earth's atmosphere, where background is defined as the preindustrial Holocene atmosphere.

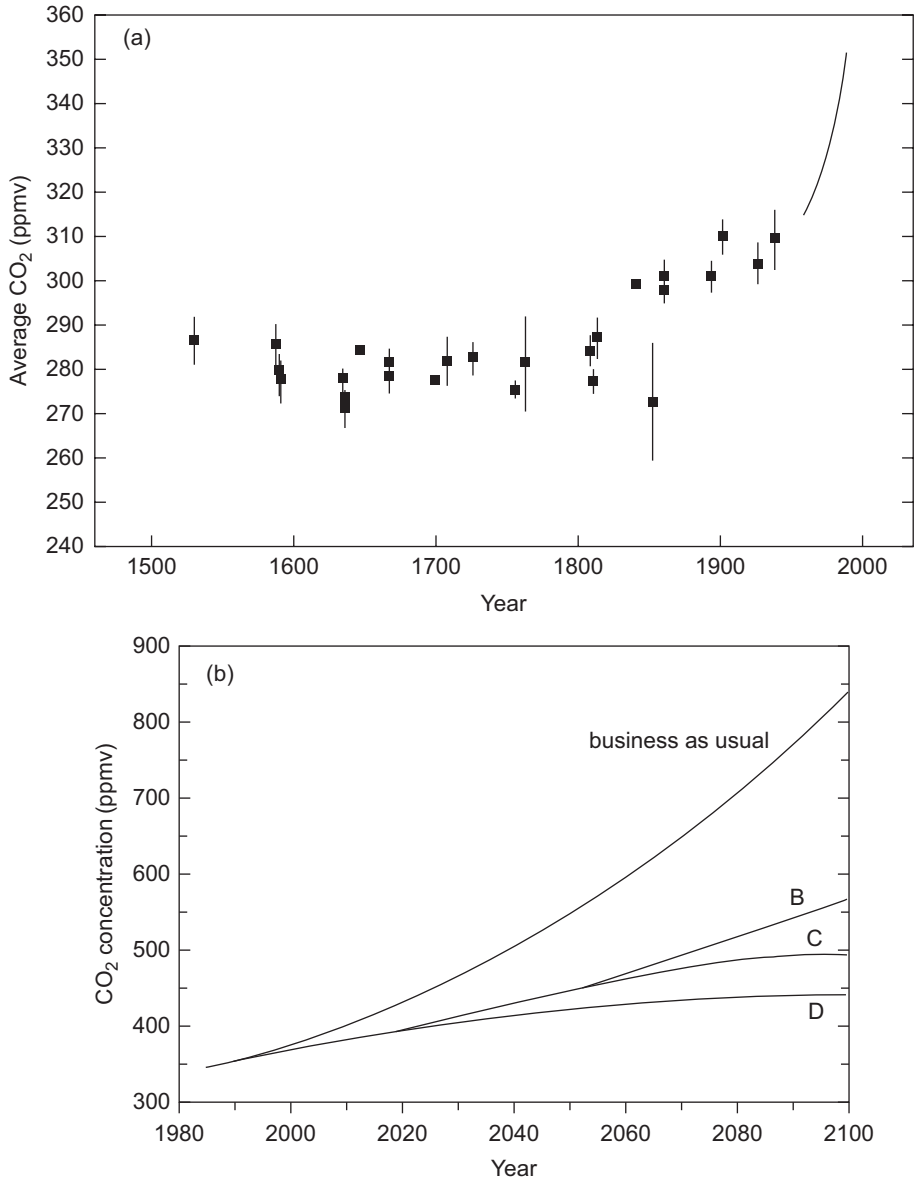


Figure 22.1 (a) Carbon dioxide concentrations in Earth's atmosphere over time from the European Middle Ages to the present. Concentration is expressed in parts per million; hence, 1 ppmv represents 0.0001%, or 10^{-6}, of the total air. (b) Projected increase in carbon dioxide levels, beginning from the mid-1980s, neglecting uptake by the ocean or continental biomass, for four possible cases described in the text. Modified from Mortensen (1996).

Projections for the future increase of CO_2 are also shown in Figure 22.1. Four cases are considered. The baseline "business as usual" assumes no change in world dependance on fossil fuels while economies and population continue to grow, at least through the first half of the twenty-first century; the current rate of worldwide deforestation also is assumed. Case B is obtained by a shift toward fuels with higher energy output per unit carbon dioxide produced, i.e., natural gas (see Chapter 23), along with cessation of deforestation and imposition of tight emission controls. Case C assumes that renewable energy sources (solar) and nuclear power take over from much of the fossil fuel use during the second half of the twenty-first century. Case D is the result of such a shift in the *first* half of that century, so that industrialized countries experience no growth in their emission of carbon dioxide. Although some uptake by oceans and continental biomass is expected, such buffers do not depress completely the atmospheric rise in CO_2, and are only temporary in any event. With a mean annual increase of 2 ppm per year over the past decade, carbon dioxide will reach 500 ppm by the end of the twenty-fist century unless, as in cases C and D, dramatic shifts are made in types of energy sources used by humans (Chapter 23).

In the rest of the chapter, we examine the physical links between such atmospheric chemical changes and alterations to the overall thermal balance of Earth's surface and atmosphere. Unfortunately, a record of temperature detailed enough in space and time to document possible changes extends no further back than about 1850; proxy records of temperature must be relied

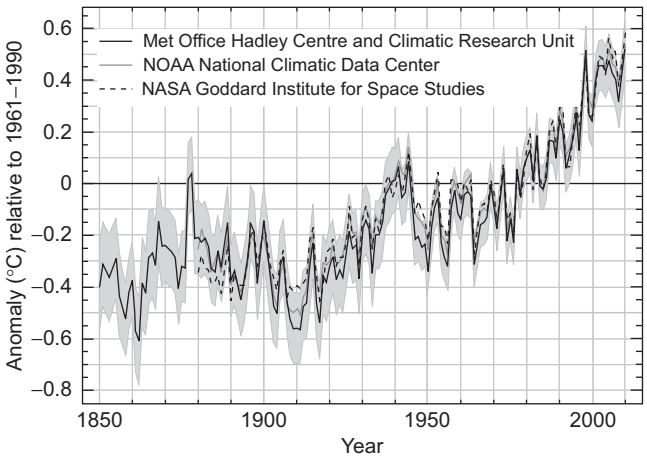


Figure 22.2 Observed annual global average temperature of Earth's surface from 1850 to the present, with the temperature change expressed relative to the average for 1961–1990. Three different sources of data and analyses are used: the United Kingdom Meteorology Office, the US National Oceanic and Atmospheric Administration, and NASA. The gray area is the 95% confidence interval, meaning statistically that there is only one chance in 20 that the actual temperature is outside that range. From WMO (2011b) by permission.

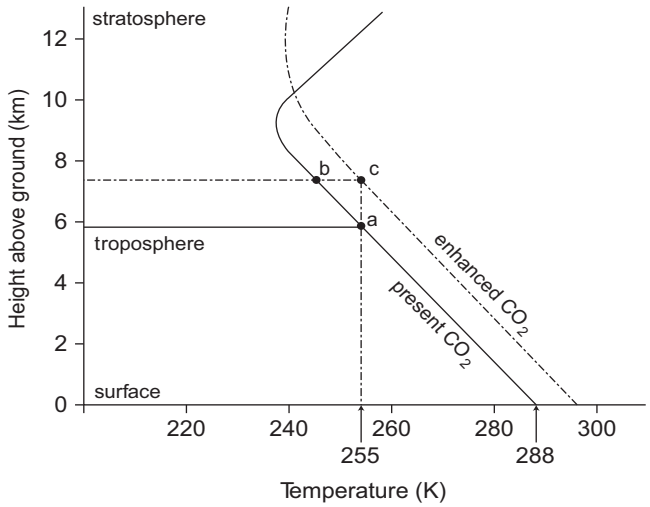


Figure 22.3 Schematic illustration of how increasing the amount of greenhouse gas in Earth's atmosphere can increase the surface temperature. The solid line represents the present temperature profile in Earth's atmosphere, and the mean radiating level (point "a") is shown. Increasing greenhouse gas concentration makes the atmosphere more opaque to infrared photons, forcing the mean radiating level upward to an altitude (point "b") where the temperature is lower. To rid the atmosphere of the same amount of heat, the temperature at the new mean radiating level must increase to point "c," forcing the whole temperature profile to increase (curve labeled "enhanced CO_2"). Modified from Mitchell (1989).

upon in spite of their less definitive nature. Even direct temperature measurements have their limitations; oceanic and continental stations have moved over time, and the expansion of cities often creates localized warmings around weather stations associated with increased concrete and less vegetation. In a number of cities, meteorological stations have had to be moved because jumps in temperature were found to be associated with building of structures and removal of grassy areas around the original stations – the so-called "urban heat island effect."

Figure 22.2 is a record of Earth's global surface temperature since 1850, for the northern and southern hemispheres combined, averaged over all seasons, from a number of continental and oceanic stations. The temperature is obtained by averaging records from these stations over the entire year. Although there are dips and plateaus in the curve, overall the climate has warmed during this time period. After 1970, temperatures began to climb and continue to do so with only small hiatuses. Global temperatures in the past two decades exceed those in the nineteenth century by nearly 1°C. Because the rise encompasses both hemispheres in a record obtained over a large number of stations, it cannot be primarily a result of the urban heat island effect.

A check on the direction and the magnitude of the rise comes from studies of valley glaciers around the world, essentially all of which have retreated up the valley hundreds of meters or kilometers since the middle-1800s. Careful measurement of the retreats, combined with models of how much temperature increase is required to produce a given amount of retreat, allows an estimate of the past century's worldwide temperature increase independent of weather stations. Such glacial studies by Dutch climatologist J. Oerlemans indicate a worldwide temperature rise of roughly 0.7°C, with an estimated error of plus or minus 0.2°C. This number is close to, and consistent with, the globally averaged temperature rise derived from weather stations.

Although the global mean temperature of the last few years is at least as warm as any in the past 500 years, it is still not the warmest in the Holocene: the Holocene Climate Optimum of 5,000 to 9,000 years before present appears to have been hotter based on ice core, sediment, and other data. Furthermore, we do not know whether the global average temperature will remain constant, fall, or rise further in the coming decades. However, physical understanding of the greenhouse effect by which Earth's climate is maintained above the water freezing point provides a very strong argument in favor of the notion that much if not all of the temperature increase is due to the enhanced flux of greenhouse gases into the atmosphere caused by human activities.

22.2 Modeling the response of Earth to increasing amounts of greenhouse gases

22.2.1 Review of basic greenhouse physics

The basic physics of the greenhouse effect was described in Chapter 14. As the amount of infrared-absorbing gases is increased, Earth's atmosphere becomes more opaque to infrared photons. The altitude above the surface at which such photons are finally free to escape (the *mean radiating level*) therefore moves upward, toward lower air density, as greenhouse gas concentration increases. However, as Figure 22.3 shows, because

the temperature falls with altitude in the troposphere, the new mean radiating level is colder than the old one, and hence less efficient at removing energy. Its temperature must increase, raising the entire temperature profile of the troposphere. In this way, increasing greenhouse gas concentrations raise the mean surface temperature of Earth.

This effect can be expressed in terms of the "mean radiative forcing" of greenhouse gases, which is the change due to greenhouse gases of the net irradiance (solar photons minus infrared photons) the atmosphere radiates as measured at tropopause. While this sounds complicated, it effectively means the added power associated with more greenhouse gases, which cause the atmosphere to absorb more of the infrared energy moving upward from the surface. The radiative forcing thanks to all greenhouse gases has risen by 2.8 watts per square meter relative to what is calculated for 1750, and by 1.1 watts per square meter relative to what was measured in 1980. While this seems small relative to the total solar input of roughly 1,300 watts per square meter, remember that the atmospheric temperatures are the result of a balance between incoming and outgoing solar radiation, and thus small changes in the amount of infrared radiation the atmosphere absorbs have a big effect. Further, note that 40% of the increase in the mean radiative forcing since 1750 has occurred in just the last 30 years (1980–2010).

The basic physics of the greenhouse process described above is straightforward enough that there is little argument about its validity. We know, for example, that Earth's neighboring planet, Venus, receives less sunlight at and near its surface than does Earth because of a layer of bright, reflecting clouds. However, the surface temperature of Venus is over twice that of Earth's, and the atmosphere is possessed of a carbon dioxide pressure of 90 bars, 300,000 times the amount of CO_2 in our atmosphere. It is not too great a leap to infer that the Venusian atmosphere is in a state in which the enormous amounts of carbon dioxide create a greenhouse effect much larger than Earth's, and models show that the surface temperature and CO_2 abundance are indeed consistent with each other.

22.2.2 Some complications

As simple as the basic physical concept is, it does not fully describe the actual situation. The most fundamental complication is that water vapor is also a greenhouse gas, but its abundance in the atmosphere depends on the global mean surface temperature through evaporation from the ocean and rainout in precipitation events on land, ice, and sea. Cloud formation (see below) complicates any direct relationship between increases in surface temperature and consequent increase in water abundance. However, it appears that as other greenhouse gases increase the global average temperature, a slight positive feedback occurs through an increase in the amount of atmospheric water vapor.

Another complication is that radiation (transport of photons) is not the sole means of the movement of heat energy outward. Particularly in the lower part of the atmosphere, the temperature profile becomes so steep (decreases so sharply with altitude) that bulk air movement (that is, convection) plays an important role. *Dry convection* involves bulk movement of air without condensation of water to form clouds; it occurs in the lowermost part of the atmosphere and particularly in dry regions. *Moist convection* includes the effects of cloud condensation and evaporation, which add and delete heat from the surrounding air. Cloud formation most often is a result of air containing water vapor rising, expanding as the surrounding pressure drops with altitude, and then cooling until the air can no longer hold the water as a gas. The *dew point* is thus reached, and water condenses out to form small liquid drops or solid ice particles.

Moist convection is a sufficiently energetic process that it alters the environment around it and the consequent transport of energy. Large amounts of water in an atmosphere initially unstable (tending toward bulk air motions to remove heat) can create large thunderstorm complexes, in which updrafts and downdrafts may reach all the way up to and beyond the tropopause (defined in Chapter 15). This is particularly the case in the tropics, but large storm complexes also dominate weather in midlatitude continental regions. The convective transport of heat, particularly involving moist convection, alters the relationship between greenhouse gas increase and the temperature response of the atmosphere; by how much (and even in what direction) remains a matter of dispute.

Formation of clouds also alters the radiative balance of the atmosphere, aside from the effects of moist convection. Clouds can reflect, scatter, and even absorb incoming solar visible radiation; they also may absorb infrared radiation moving upward from the deeper atmosphere. The overall effect of clouds on global climate is complicated. The difficulty arises from the wide range in shapes of clouds, size of the cloud droplets or ice particles, the breadth of altitudes over which clouds form and extend, and conditions under which precipitation (rain, hail, sleet, or snow) forms. Some cloud types may lead to a net warming of the atmosphere, whereas others will cool it. Hence, if global temperatures increase because of enhanced greenhouse gases, and the resulting increased moisture (from more vigorous evaporation of ocean water) creates more cloudiness, the net effect of that cloudiness depends largely on the types of clouds and their mean altitude. Recent satellite and aircraft measurements of the amount of visible and infrared radiation coming out of, and moved around within, clouds are beginning to untangle these very complicated effects.

Snow, continental ice sheets, and sea ice provide very highly reflective surfaces that prevent much sunlight from being absorbed at high latitudes on Earth's surface. As global temperatures increase, the amounts of land and sea ice and snow will decrease, causing more sunlight to be absorbed and amplifying the greenhouse warming. How much of an amplification will occur depends on the details of the response of the ice and snow to warming. Increased precipitation at high latitude, another likely result of warming, could actually increase snow and ice cover in winter at high latitudes and/or altitudes, providing a moderating effect to the amplification.

Variability in the output of the Sun affects the amount of energy the atmosphere must transport back out, and has the potential to obscure the signature of human-induced global warming. Measurements of the Sun's luminosity taken over the past couple of decades show that it has varied only by plus or minus 0.02%. Compared to the effects of increased CO_2 over the same period, this number is quite small and, although some climate amplifications of the solar variability are possible, they

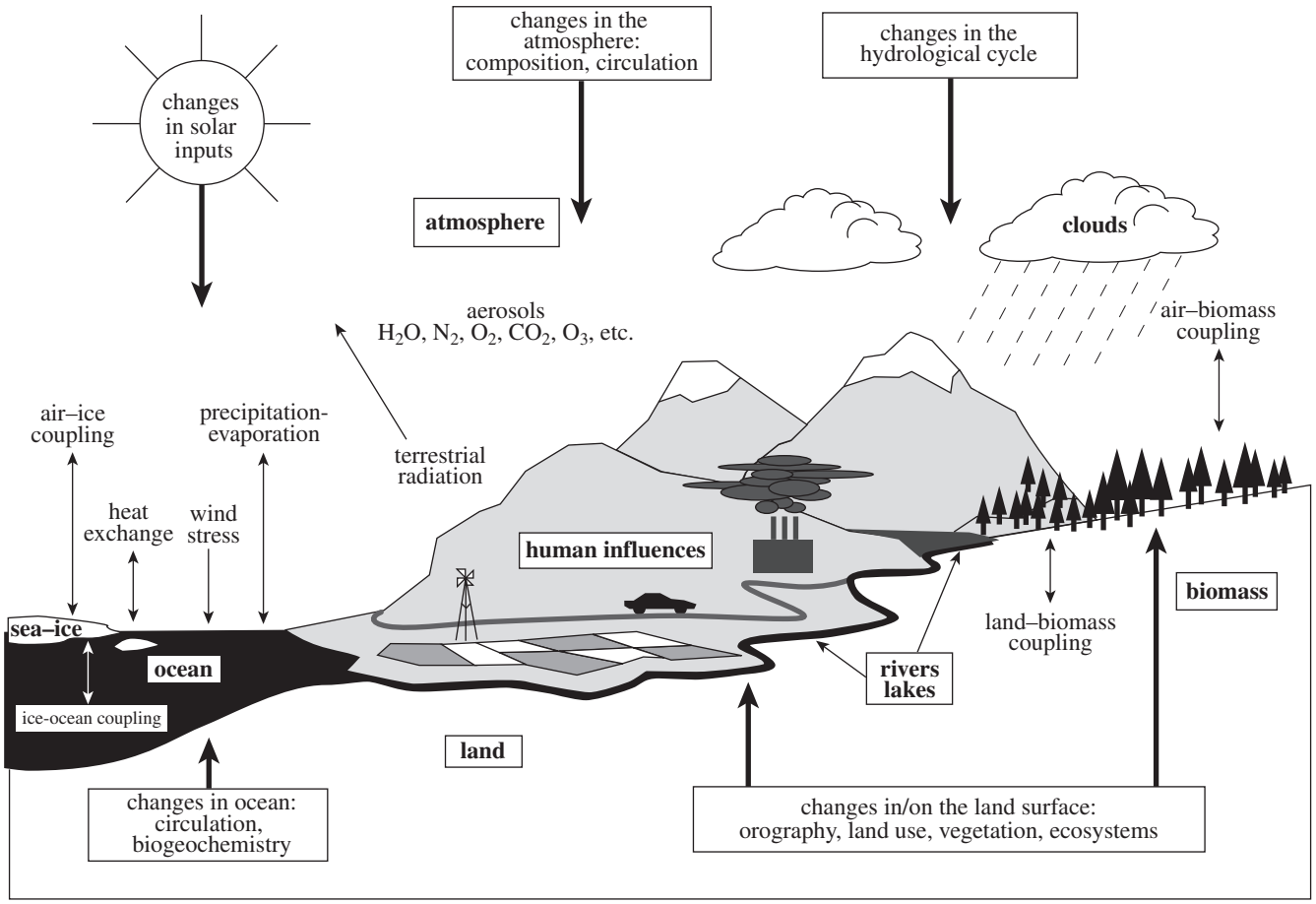


Figure 22.4 Processes affecting the nature of climate today, with an emphasis on the changes that might result from human influences. *Wind stress* is the movement of ocean water caused by the action of wind; *biomass* refers to biological organisms both living and dead, that interact chemically with the atmosphere, land, and oceans. From Trenberth *et al*. (1996).

are unlikely to reverse or dominate global warming associated with increasing carbon dioxide. On longer timescales, the Sun's luminosity varies more significantly (Chapter 14), but projections of human-induced global warming are concerned primarily with the next half-century, a time not much longer than that over which detailed solar measurements have been made.

Perhaps the most important uncertainty lies in the role of the oceans. A thorough discussion of this is deferred to section 22.5, because of its complexity. Figure 22.4 illustrates graphically how the processes discussed above fit together and emphasizes that climate is not simply a matter of the vertical structure of the atmosphere, but also of what is happening from place to place on Earth's surface and in its oceans. We know from our experience with weather patterns that the three-dimensional nature of the planet is important. To capture this aspect of the problem requires rather involved computer models, to which we now turn.

22.2.3 General circulation models

One-dimensional climate models simulate the transfer of energy and matter only in one direction, namely, up and down. However, on a planet, energy and matter also move sideways in the atmosphere and on the surface. It is useful to define the sideways direction parallel to a line of latitude as *zonal*, and parallel to a line of longitude as *meridional*. Because Earth is roughly spherical, different latitudes receive varying amounts of sunlight; even though Earth's axis is tilted, the equator receives the largest amount of heat averaged over the year. As a consequence, heat tends to be redistributed by the oceans and the atmosphere in a meridional direction, that is, from the equator to pole. Warm tropical air rises, moves away from the equator, and sinks; this cycle is repeated at higher latitudes.

The Earth also spins on its axis, and this spinning motion modulates the transport of heat from equator to pole. Sinking air in the northern hemisphere is forced to spin clockwise, and in the southern hemisphere counterclockwise. Regions in which air is drawn inward by low pressure, forced to rise and form clouds and precipitation, will rotate counterclockwise in the north and clockwise in the south.

These systems of high and low pressure produce much of the weather with which we are familiar at middle and low latitudes. Their sense of rotation, induced by Earth's spin, interacts with the distribution and shape of continents to produce complex patterns. Low pressure spiraling counterclockwise as it moves eastward across the central United States draws moisture off the Gulf of Mexico to produce the well-known severe thunderstorms that often plague Texas, Oklahoma, Arkansas, and other midland

states. The positions of high and low pressure systems in the Pacific and Indian Oceans determine each summer the strength of the south Asian monsoon rains, critical to food production cycles for billions of human beings.

To simulate such complex weather patterns, computer models must do more than calculate how photons are absorbed and re-emitted on their way out of the Earth's atmosphere. They must also keep track of how energy (heat), moisture, and bulk air flow from one region to another. Models that do this are called *general circulation models*, or GCMs. The strategy is to divide Earth into a checkerboard in which each square, or grid point, is as small as possible; smaller gridding requires faster computers to handle the more numerous grid points. Newton's laws of motion, along with the laws of thermodynamics, are applied to the air, water, and heat in each grid, and both matter and energy are allowed to flow from one grid point to another. Sunlight shines according to latitude, season, time of day, and amount of cloudiness. In this way the flow of moisture, winds, and heat around Earth can be simulated on large, fast computers. Such GCMs, relying on detailed temperature, wind, pressure, and moisture information from thousands of weather stations worldwide, are used to predict weather several days or more in advance. General circulation models have also been adapted to predict atmospheric circulation patterns on other planets, as well as the nature of the climate at earlier times in Earth's history. They are the basic computational tool for evaluating climate change caused by increasing abundance of greenhouse gases.

As carefully constructed as they are, GCMs have limitations. The first of these is intrinsic to the nature of climate itself. The ocean, atmosphere, and land form a coupled, *nonlinear* physical system. In recent decades the properties of such systems have been investigated and found to exhibit chaos. Chaos does not imply complete randomness (Chapter 19). However, such systems can evolve into many different states, unlike simpler systems. A simple system, started out in two slightly different configurations, will diverge rather slowly in appearance. A chaotic system, started out in two different states, will *exponentially* diverge in its characteristics – an almost explosive parting of the ways between the two slightly different starting states. Insight into the difference between a simple and a chaotic system is not easy to gain, but Figure 3.3 may be of help: it shows that an exponential function of a parameter always grows more rapidly than a power-law function of the same parameter. The general nature of a chaotic system can be described from a probabilistic point of view, but not its details. Climate has this nature. Thus, although GCMs are very good at using data to predict trends in climate over various periods of time (months, years, decades), they cannot completely capture the details of climate fluctuations (which we perceive as "weather"), and may fail to identify when Earth's climate could shift into a drastically different state.

The second limitation has to do with grid size. Most GCMs today are limited to grid sizes of a hundred kilometers in each direction. Smaller grid sizes come at the cost of vastly increased computing time to obtain a result. However, weather is affected by processes on much smaller scales; mountains, shapes of coastlines, and changing surface characteristics may occur on scales of ten kilometers or less. Moist convection cloud and rain formation must be characterized on kilometer and smaller scales. These *subgrid* processes play key roles in determining the movement of air, moisture, and heat around Earth, yet they cannot be explicitly computed in GCM models. The strategy is to try to approximately characterize such processes computationally so that, on the scale of a grid point, they produce the same effects that the real processes do. Studies of how well GCMs account for the effects of moist convection, for example, suggest that, as yet, this strategy is only partly successful.

The third limitation of general circulation models lies in the coupling of atmospheric and oceanic processes. Because the nature and causes of ocean circulation patterns are only incompletely understood, no model exists today that fully characterizes how the atmosphere and the ocean interact. General circulation models may be particularly sensitive to this limitation because of their large demand for computing power and the difficulty of handling simultaneously the short timescales of the atmosphere (days) and the long timescales associated with ocean mixing (centuries). However, much effort over the past decade has been put into improvements in the accuracy of the air–ocean interaction in such models, based on better understanding of oceanic circulation, the detailed physics of the exchange of material between ocean and air at the sea surface, and increased computing power. The most recent GCMs, to emphasize their more sophisticated incorporation of coupled ocean and atmosphere processes, are sometimes called atmosphere–ocean global circulation models (AOGCMs).

AOGCMs represent the most detailed and accurate simulations of Earth's climate that is available with present-day computing power. As computers continue to improve in speed and memory, the challenge will be to incorporate physical processes with greater fidelity. It is part of the nature of scientific research to test models of physical systems against their real behavior, based on observational data. With expanded means of collecting data on the current Earth environment, as well as on those of other planets and the Earth in its past, the reasonable expectation is that AOGCMs will continue to improve in their capability to elucidate the behavior of Earth's climate and make forecasts of future climate changes. By way of example, Figures 22.5 and 22.6 illustrate climate-change predictions of some state-of-the-art GCMs.

22.3 Predicted effects of global warming

The inherent and unavoidable uncertainties in the output of scientific models such as GCMs have led to confusion among many people about the validity of global warming and, more specifically, about the human-induced component. Colloquially, when we say something is "uncertain," we mean that it may or may not happen – or may or may not be true. "I am uncertain as to whether Jill was accepted to Harvard" means that Jill might have been admitted to Harvard, or might not. It is a yes-or-no proposition, with the only solution being to ask Jill (and then it is possible she might lie about it). In science, uncertainty has a different meaning: it refers to the dispersion of values attributed to a measured quantity or numerical output of a computer model. The conclusion of an experiment may be definitive even when measurement uncertainty – random errors – remain present in the output data. Indeed, no experiment or numerical model can

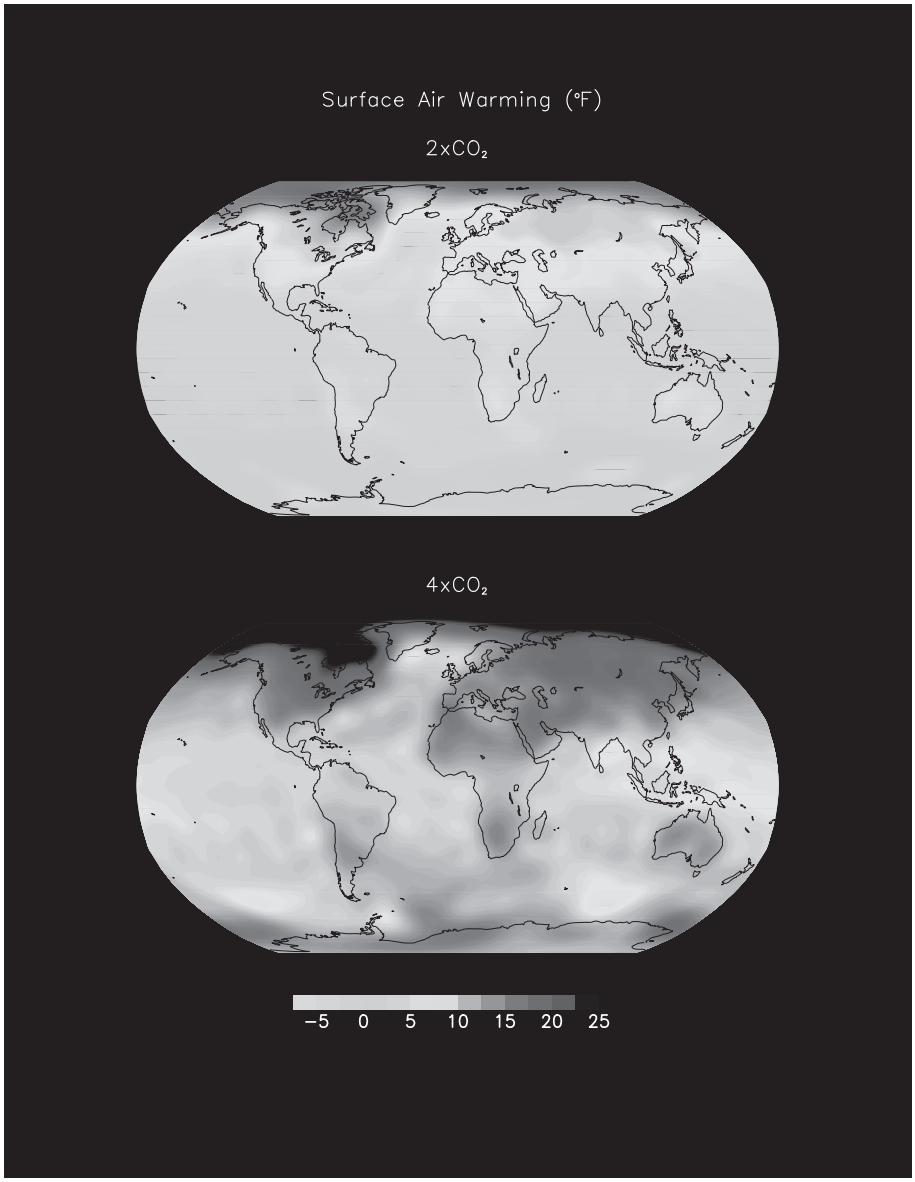


Figure 22.5 Predicted rise in surface air temperature for CO_2 doubled and quadrupled, based on a climate model developed at the Geophysical Fluid Dynamics Laboratory of Princeton University. Temperature is shown in degrees Fahrenheit. Figure created by Thomas Knutson, provided courtesy of the Geophysical Fluid Dynamics Laboratory, National Oceanic and Atmospheric Administration. See color version in plates section.

be without uncertainty, but this does not mean that the outcome is not known or understood. If you want to move a piano through a doorway, you will need to measure the width of the doorway and of the piano. Measure each one ten times and you will get ten different answers for each. But, as long as you consider the range of measurement errors – the uncertainties – in assessing the width of each – you will be able to come to a definitive conclusion as to whether you can get the piano through the doorway.

In climate science, one must be careful to distinguish robust models and data from models that – either because they seek to predict more complex phenomena or require data that are lacking – carry significant risk of not making correct predictions. The basic relationship between overall surface temperatures and abundances of greenhouse gases is understood in a robust fashion and the data on greenhouse gas abundances and temperatures – averaged over many stations – are of high quality and low uncertainty. That the atmospheric radiative balance is changing, and hence Earth's globally average surface temperature is increasing, in response to greenhouse gases introduced by human activities is a robust result that no longer has a plausible counterargument. Even the magnitude of the effect, which was highly uncertain a few years ago, seems well understood. However, the specific effects of the change in the atmosphere on regional weather patterns, rainfall, severe storm events, etc., are much more difficult to quantify with confidence because of the complexity of the models and the chaotic nature of weather events.

Committees of scientists are convened at the behest of governmental agencies to assess the work of various groups in global climate change research, and to make a consensus determination as to which predictions of climate modeling are reliable

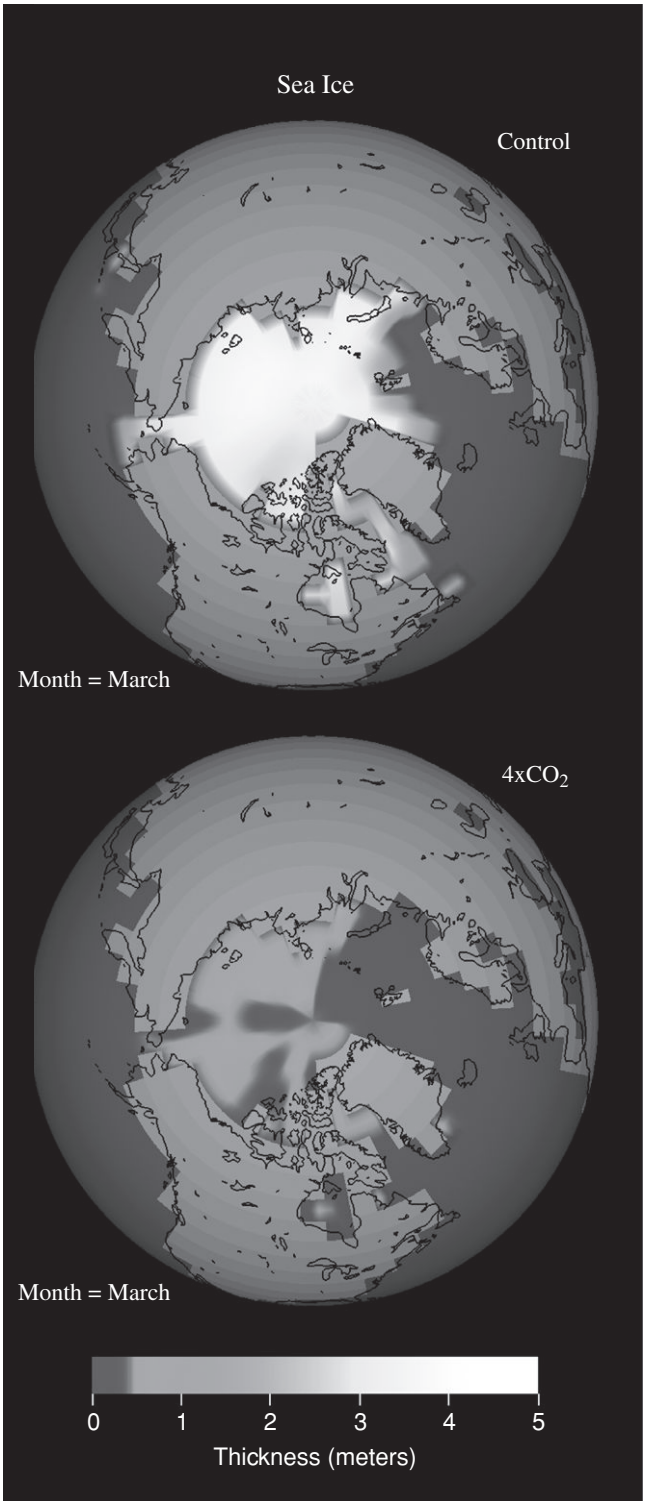


Figure 22.6 Changes in sea ice thickness for quadrupled CO_2 (bottom) relative to no change in carbon dioxide (top panel). Figure created by Hans Vahlenkamp and Thomas Knutson, provided courtesy of Geophysical Fluid Dynamics Laboratory, National Oceanic and Atmospheric Administration. See color version in plates section.

as opposed to speculative. The International Intergovernmental Panel on Climate Change (IPCC) has served in this capacity since the late 1980s, and shared a Nobel Peace Prize for their efforts. Other national and international bodies such as the US National Research Council have conducted similar assessments. The results of these deliberations serve as a useful summary of potential effects of the human-induced increase in greenhouse gases, by providing an ordering of effects from most to least probable. Some of the potential impacts besides the increase in global mean surface temperature are summarized below, in rough order of decreasing certainty in their occurrence.

22.3.1 Large stratospheric cooling

The stratosphere, the region above 10- to 15-km altitude, is an important contributor to the heat balance of the atmosphere. Because the air is so thin at those altitudes, infrared photons at many wavelengths are free to move upward and out to space. Temperatures in this part of the atmosphere increase steeply with altitude, however, because ozone (O_3) and other gases can absorb ultraviolet photons from the Sun. As shown in Figure 22.3, the increase in infrared opacity of the lower atmosphere shifts the minimum temperature point (tropopause) upward; in effect, the level at which infrared photons become capable of escaping moves upward. The stratospheric increase of temperature with altitude thus is delayed to higher levels, and the net result is a cooling of much of the stratosphere. The importance of this cooling is that it provides a test of a fundamental aspect of the greenhouse atmospheres. If stratospheric cooling were not occurring, basic aspects of the theory might be wrong. Complicating the signature of the cooling are the depletion of ultraviolet-absorbing ozone from introduction of industrial CFCs into the atmosphere, and the warming effect of sulfate aerosols that are injected into the stratosphere by volcanic eruptions. The cooling effect of decreased ozone may, in fact, dominate over the effect of increased carbon dioxide. Satellite and balloon data show that stratospheric temperatures have been decreasing since 1979, but sudden warmings due to volcanic eruptions are apparent as well.

22.3.2 Global mean increase in precipitation

Increased sea surface temperature means increased evaporation rate over the oceans, leading to an increase in precipitation averaged over the globe. However, the distribution of this increased precipitation over the globe will be highly variable and may be difficult to predict accurately with current climate models, primarily because of the large area covered by each grid point. Furthermore, although precipitation may increase, many continental areas will have drier soils because the higher temperatures also will increase local continental evaporation rates. Models suggest that, in most locations, the increased precipitation will not compensate for this increase of evaporation, and desertification (conversion to a more arid regime) might result over food-producing areas.

22.3.3 Northern polar winter surface warming

Evidence from studies of the Cretaceous and other paleoclimates described in Chapter 19 suggest that the polar regions of Earth

experience amplified climate change relative to lower latitudes. All current model simulations show a maximum surface warming in high northern latitudes in winter, but much less arctic warming in summer. It is thus very probable that sea ice in the northern hemisphere (where no polar continent exists) will be reduced substantially in extent in the coming decades. Predictions are less certain for the southern hemisphere because of the complicated pattern of sea ice and ice shelves associated with Antarctica.

22.3.4 Rise in global mean sea level

The sea will rise because of the expansion of the water as it is warmed. This is the most certain but by no means the most important component of sea-level rise. Shrinking ice caps, continental glaciers, and the major Greenland and Antarctic ice sheets may play important but less certain roles. If the trend over the past century were to continue, an additional half-meter of sea-level rise would occur between now and 2100. Folding in the possible accelerated melting of ice sheets due to global warming effects leads to a less certain, but larger, value up to one meter by the year 2100. An even larger value could obtain if the West Antarctic ice sheet collapses, but this is a highly speculative possibility. The effects of sea-level increases are somewhat controversial and complex, because much of the damage occurs during storms that locally raise sea level due to the low pressure and effects of onshore winds. However, given the large number of human habitations in coastal areas, essentially at sea level, increased disruptions and property damage are all but inevitable.

22.3.5 Summer continental warming and increased dryness

This category and the ones below are all much less certain that those above, because of the dependence on details of physical processes that are complicated and not yet fully simulated by computer models. As noted above, increased precipitation worldwide does not mean wetter continents. Most computer models show a decrease of soil moisture in the interior mid-latitudes of southern Europe and North America during the summer. This is mostly due to earlier springtime melting of snow in the warmer mid- and high-latitude springtime, higher summertime temperatures leading to increased evaporation from soils, and (in some regions) decreased summer rainfall.

22.3.6 Regional vegetation changes

Changes in temperature, rainfall, and soil moisture patterns will drive species poleward or to higher elevations. The extent of the process is uncertain, but the rapidity of the global warming compared to natural climate change will cause major ecological disruptions and possible species extinctions. Preserves set up today for endangered plant and animal species will be abandoned by those species as they move poleward, possibly being interrupted in their retreat by major urban or developed areas. Additionally, in mountainous country, species will be forced to move upward in response to increased warming and drying; uppermost mountain ecosystems may become locally extinct. The process of such extinctions may not be a gradual one: unusually hot and dry summers in the American Southwest contributed from the mid-1990s through 2010 to devastating mountain forest fires and consequent shrinkage in the distribution of rare species.

22.3.7 More severe precipitation events

The ability to predict the nature and distribution of precipitation events using GCMs is limited. Warmer temperatures lead to a more vigorous cycle of water evaporation and rainfall, leading to more severe droughts or floods in some regions. Several models show an increase in intensity of rainfall or snowfall on a warming globe, suggesting the possibility of an increased frequency of extreme precipitation events. A warmer, wetter atmosphere sitting atop a warmer ocean might trigger more frequent tropical storms and hurricanes, as well as increase their intensity. Because such storms are affected by many factors, however, the extent of this effect is uncertain; what is known is that such storms are formed today over warmer ocean waters and lose their strength over cooler surfaces.

22.3.8 Changes in climate variability

As global temperatures warm, the response of the oceanic–atmospheric system may change in unpredictable ways. Because of the rapid input of greenhouse gases into the atmosphere, and of surface changes in forest cover, the complex climate system of the Earth may try to shift unexpectedly into different states, on short (decadal) timescales. Evidence for this comes from the surprisingly rapid changes seen in the ice-core climate record at the boundaries between glacial and interglacial episodes.

22.3.9 Regional-scale changes will look very different from the global average, but their nature is uncertain

Prediction of how climates will change within distinct regions of hundreds of kilometers extent is less reliable than predictions of globally averaged changes, but will improve as more powerful computers allow finer grid sizes in AOGCMs. How worldwide climate feeds into changes in localized deserts, mountain regions, grain belts, etc., is crucial to forecasting the economic impacts of human-accelerated climate change.

22.3.10 Biosphere–climate feedbacks

In contrast to the interaction of ocean and atmosphere, interaction between the biosphere and the climate system remains poorly characterized. In particular, how both oceanic and continental photosynthesizers will react to enhanced levels of CO_2, as well as warming, remains a matter of debate. Experiments involving subjection of plants to enhanced levels of carbon dioxide show both positive and negative effects, dependent on the species of plant, the amount of increase in carbon dioxide, and the protocol of the experiment. Even if ongoing experiments provide a complete understanding of the physiology, the large-scale interaction among oceans, atmosphere, and organisms will remain a challenge to quantify.

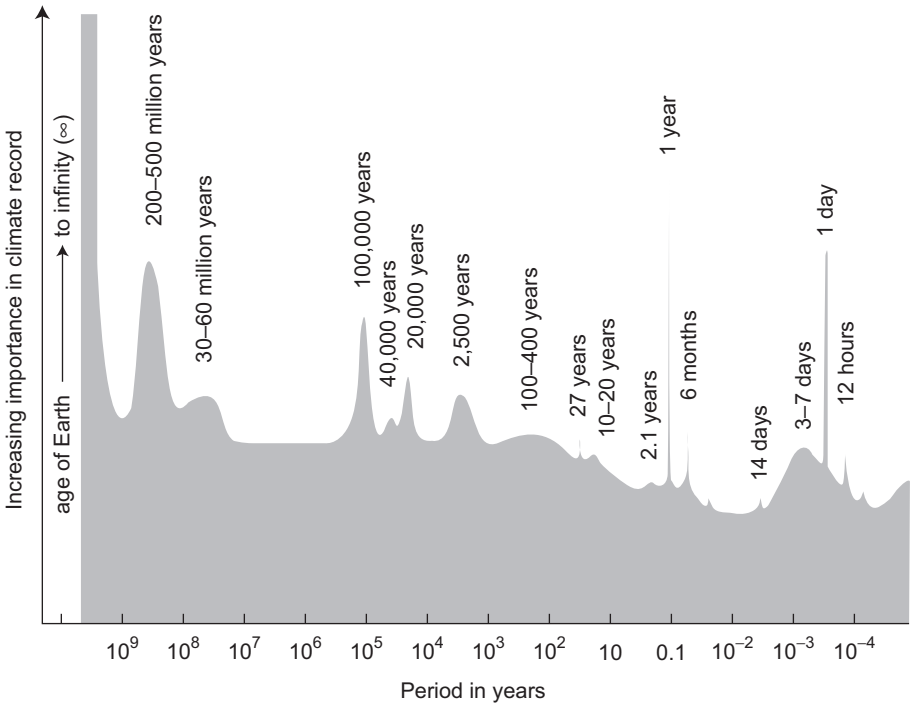


Figure 22.7 Timescales over which climate changes. The graph is a model based on a variety of data revealing climate change. Shown is the "importance" of the temperature variation, averaged over Earth, as a function of timescale. Where Earth's climate shows a tendency to fluctuate or change on a particular timescale, the value of the temperature importance in the climate record will be high. Because of Earth's rotation, daily and semidaily variations in temperature are a very strong component of climate, as are seasonal effects. There are variations also on decadal timescales, having to do with the pattern and distribution of droughts. The Little Ice Age shows up as a bump on the timescale of several centuries. Longer term climate changes that show up on the graph include those due to the Milankovitch variations in the orbit of Earth, and those over millions of years caused by the movement of the continents. Modified from Peixoto and Oort (1992) by permission of Springer-Verlag.

22.3.11 Details of life in the next quarter century

Because of the oceans' capability for temporarily storing heat and carbon dioxide, we have already bought into a significant amount of global warming over the next 25 years, associated with the increase to date in carbon dioxide – even if we stopped producing greenhouse gases today. Possible effects outlined above have human implications, of which some are:

1. lower food production due to loss of agricultural land
2. increased damage or loss of coastal areas due to high sea level and storms
3. loss of treasured ecological preserves, recreational areas, and species due to rapid climate change
4. increased discomfort and energy consumption due to more frequent occurrence of extremely hot days
5. decreased health and increasing disease vulnerability due to heat stress and poorer nutrition
6. slowing of population growth, or even population decline.

22.4 The difficulty of proof: weather versus climate

How can we detect the onset of human-induced global warming? There are natural and human-induced effects that can obscure the temperature signature of increasing CO_2 and other greenhouse gases. First, the oceans tend to slow and smooth out the effects of increased greenhouse gases over many decades. Second, weather obscures climate trends. The day-to-day, season-to-season, and year-to-year variations that make up what we call weather are only part of the natural variations in climate extending from hours up to millions of years (Figure 22.7). Possible human-induced increases in global mean temperature over the next decade will be partially obscured by the inherent natural variability of Earth's climate.

Yale climatologist M. E. Mann and colleagues have assembled climate data over the past five centuries from a number of sources, including ice cores, isotopic ratios in sea corals, tree-ring records, and historical accounts. They find that climate over the past half-millennium shows particular variability on timescales of several decades, as well as on timescales of several hundred years. The shorter of these timescales is troublesome for watchers of potential global warming because reliable instrumental temperature and precipitation records are available only for the past hundred years. This is really too short a span to allow removal of a natural climate variation that might cycle once every several decades. The longer cycle of several centuries also produces ambiguities in interpreting the temperature records over the coming decades; we appear to have recently come out of a particularly cold period (the Little Ice Age), and some have argued that the bulk of the warming experienced in the twentieth century is simply the continuation of that trend.

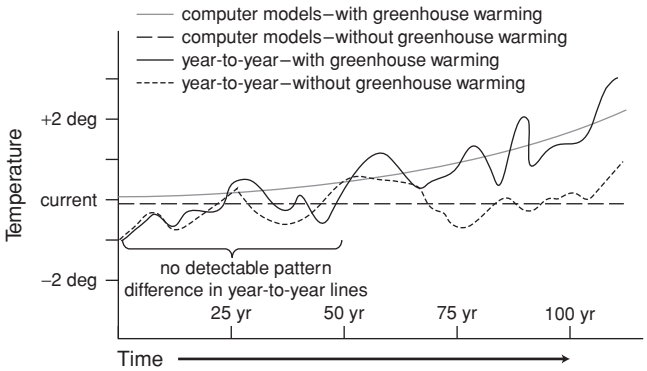


Figure 22.8 Cartoon illustrating the obscuring effect of natural year-to-year weather variations on global climate change. The graph does not contain actual data, but is intended to illustrate a concept. The lines labeled "computer models" are intended to represent global mean temperatures, in which year-to-year variations are smoothed out, for the cases of global warming and no-global warming, respectively. The "year-to-year" lines include annual and decadal variations. Such variations could prevent small amounts of global warming (be it human-induced or otherwise) from being detectable.

Of course, the climatic exit from the Little Ice Age might be the result of human-caused input of greenhouse gases into the atmosphere, which became significant in the same period that we came out of the Little Ice Age. And the extent of the warming since the late 1970s rivals anything seen in proxy climate data over many centuries, suggesting something more than just a return to baseline Holocene conditions.

Particulates in the atmosphere (aerosols) can act to obscure the signature of global warming. It has been proposed that the flattening of the temperature curve from the 1950s to the 1970s (Figure 22.2) might have been the result of an accumulation of industrial sulfate aerosols in the atmosphere, blocking some of the sunlight from reaching the surface. Tighter pollution restrictions cleaned up many sulfate-producing factories, leading to a reduction of such aerosols and, perhaps, an increase in sunlight reaching the surface. Major eruptions of volcanoes spew sulfur compounds into the stratosphere, where they condense as sulfate aerosols that block sunlight for several years until they sediment out. Such aerosols produce spectacular, red sunsets worldwide because the stratospheric winds blow the material quickly around the globe after injection into high altitudes by the eruption. Recent eruptions such as that of Mt. Pinatubo appear to have cooled the globe by a few tenths of a degree for several years. More such events, although they cannot be predicted, are to be expected. The AOGCMs incorporate the effect of aerosols, and predict less warming and less precipitation change (for a given carbon dioxide increase) than did the aerosol-free models prior to them.

Figure 22.8 illustrates the problem of extracting a trend from noisy data. The thin solid line represents what computer models predict: the slow, gradual warming of Earth due to the increased greenhouse gases. The horizontal dashed line is the average temperature that we might have in the absence of human greenhouse gas increase. Now add the weather: the short-dashed line is the temperature variation one might get year to year in a climate that does not have human-induced greenhouse warming. The thick solid line is the year-to-year global temperature for an Earth with a human-caused greenhouse warming. Given plausible year-to-year and decadal variations it takes many years to notice possible human-induced global warming, as fluctuations in temperature obscure the rise in global mean temperature. However, trends in the last decade (2000–2010), coupled with the prior record, have made it essentially certain that human activities have changed the climate on a global scale.

22.5 Role of the oceans in Earth's climate

Hundreds of times more massive than the atmosphere, the world's oceans exchange heat, moisture, and carbon dioxide with the atmosphere. Oceans play a key role in the nature of and changes in climate, but a full understanding of that role is not yet at hand. Much of the problem lies in the difficulty of observing ocean circulation patterns; unlike the atmosphere, which can be well stocked with balloons, aircraft, and other measuring devices, the deep ocean remains relatively inaccessible. Additionally, coupling ocean circulation, moisture, and heat exchange models to those of the atmosphere is not easy. Because the ocean is much more massive than the atmosphere, the timescales over which changes occur differ greatly between the two. Atmosphere–ocean global circulation models that couple the two require significantly more computing power than atmospheric GCMs alone.

As we saw with the proposed mechanism for the Younger Dryas episode, the continents play a role in ocean circulation, particularly with regard to movement of freshwater but also through modification of rain-producing weather patterns. In turn, the oceans affect continental weather, and we will explore one well-known example, the El Niño phenomenon. The ability to model these interactions, and ultimately a full coupling between ocean, atmosphere, and land, remains as yet an unachieved goal of climate studies.

22.5.1 Basics of ocean circulation

Surface oceanic circulation is well mapped (Figure 22.9), but the ocean's vertical movements are incompletely understood. The ocean circulation pattern is driven by two basically different forces, but the relative importance of the two remains controversial. *Wind stress* is the motion of air over the ocean, coupling to the surface waters by friction, which in turn transfers kinetic energy to ocean waters beneath the surface. Such effects, acting over large distances, can drive circulation patterns within the bulk of the ocean, not just at the surface. The *buoyancy force* is associated with variations in density of ocean water, which leads to rising or sinking motions that drive circulation patterns. Two sources of density variations exist in the oceans. Temperature differences lead to warm water of relatively low density, and cold water of relatively high density. Ocean water contains large amounts of dissolved salts, and these also create density differences: saltwater is denser than freshwater. Wind stress and buoyancy do not operate in isolation from each other. Wind stresses can act to remove the very uppermost layers of surface water, revealing ocean water of a different temperature beneath. In the North Atlantic, such exposed water is warm, and releases

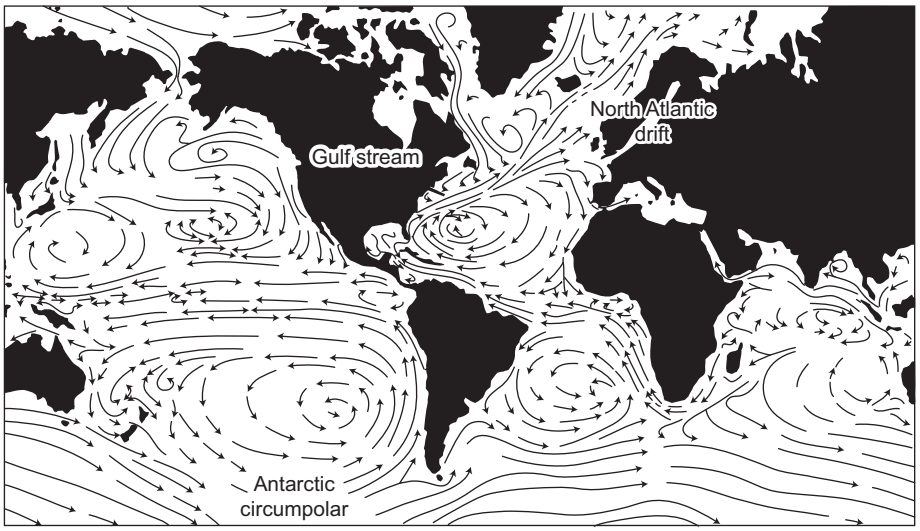


Figure 22.9 Map of ocean circulation patterns at the sea surface. Redrawn from Peixoto and Oort (1992) by permission of Springer-Verlag.

heat to the atmosphere. As the water cools it becomes denser, and sinks.

Ocean circulation is complicated by the great depth of the oceans. Deep ocean water may be isolated from surface circulation patterns on very long timescales, for hundreds or thousands of years. Such deep water may store large amounts of carbon dioxide accumulated from the atmosphere and stabilized by the large pressures near the ocean bottom. Drastic changes in ocean circulation patterns, perhaps driven by climate change, could bring deep water rapidly to the surface, where it would release this CO_2 and further perturb climate.

Ocean circulation can be tracked by sampling amounts of tritium in ocean water at various depths and locations; because tritium occurs only as a result of nuclear bomb tests beginning in the 1950s, it provides a measure of the mixing of shallow and deep ocean waters over decades since then. However, such sampling does not address the movement of the deepest and most sluggish ocean water. Instead, the content of radioactive carbon-14, which varies somewhat with time, can be sampled in deep-ocean water to determine how long it has been isolated. It appears at present that water between 1,000- and 3,500-meter depth in the Indian and Pacific Ocean represents the "oldest" deep-ocean water, i.e., that which hasn't been mixed with surface waters for the longest period of time.

The specific mechanisms (wind stress or buoyancy) by which different parts of the ocean are mixed remains in dispute. Recent models suggest that the ocean current around the broad-latitude circle surrounding the Antarctic is driven by wind stresses. Upwelling occurs in the so-called Drake passage region, and this upwelling is linked with very distant return circulations in the North Atlantic where new deep water is formed by sinking motions. Recalling that this sinking is accompanied by release of heat from shallow Atlantic waters, which plays a major moderating effect on climate there, it becomes clear that worldwide connections in ocean currents drive or modify important features of the present-day climate.

The issue of what drives ocean currents is crucial to an understanding of the role they play in climate change. Shifts in weather patterns alter both winds and precipitation. Wind shifts will change the ocean wind stress pattern, and precipitation changes can drive changes in ocean salinity and hence buoyancy. The details of these changes, the consequent response of ocean circulation, and the resulting alteration of climate remain challenging and unsolved (but important) problems.

One aspect of oceanic–atmospheric interactions involving changes in sea surface temperatures has received much attention because of its effects on worldwide weather patterns on several-year cycles. Examination of this, El Niño, phenomenon is worthwhile because it illustrates the complexity of oceanic–atmospheric interactions.

22.5.2 El Niño phenomenon

On intervals of roughly 2 to 10 years, an anomalously strong ocean current pushes unusually warm and fresh surface waters against the western coast of South America, where it moves southward carrying tropical fauna. This phenomenon is referred to as *El Niño*, a co-option of a traditional term used by Peruvian fisherman to describe the annual warming of offshore waters right after Christmas (hence, the reference to "the Child," i.e., the Christ Child of the Christian religion). Meteorologists now use the term to refer only to the unusually intense warmings.

Accompanying the warming of eastern Pacific waters is an associated warming of air temperatures. A strong El Niño may persist for more than a year. Furthermore, strong episodes involve perturbations to Pacific Ocean currents over many thousands of miles to the west of the South American coast. The atmospheric disturbance is, if anything, more widespread. During El Niño years, severe droughts occur in Australia and northern Brazil, and unusually heavy rains may occur in Ecuador and northern Peru, among other parts of the world. The Spanish conquistadors were aware of these *años de abundancia* (years of abundance) in the western parts of then-new Spain but did not connect them with changes in ocean currents. In fact, the alteration in weather patterns had been understood as a separate phenomenon involving shifts in positions of seasonal high

and low pressure patterns, known as the *Southern Oscillation*. It was not until the 1960s that the link between it and El Niño was recognized. This link is now accepted and the combined phenomenon is referred to somewhat clinically as *ENSO* (El Niño/Southern Oscillation).

The mechanics of the El Niño phenomenon involve waves generated in the Pacific Ocean that push warm water toward the east, raising sea level there slightly and depleting the amount of warm water in the western Pacific. The onset and dissipation of this event involve a complex, not wholly understood, series of feedbacks between the changed sea-surface temperature distribution and the atmospheric temperature pattern. What is not understood at all is the origin of the ENSO phenomenon in its entirety. Proposals range from the idea that the ocean circulation pattern itself drives the oscillation, all the way to the somewhat exotic notion that undersea eruptions contribute heat to the deep ocean and stimulate the effect. Also poorly understood is whether oceanic–atmospheric oscillations comparable to those in the equatorial Pacific might exist elsewhere. Finally, the speculation has been made that ENSO represents the signpost of a relatively warm climate on the edge of instability, and that not all times during the Holocene experienced such oscillations.

Much research is devoted to understanding all of the oceanic and atmospheric connections associated with ENSO, as well as what its true recurrence pattern (cyclical or chaotic) is. These connections may be subtle. For example, one proposed effect of ENSO is a temporary slowing in the growth of atmospheric CO_2 abundance, through an alteration in ocean circulation that inhibits the usual movement of CO_2-rich deep waters to the surface. Larger shifts in ocean circulation caused by processes other than ENSO have the potential for more drastic changes in carbon dioxide abundance and hence climate, but on long timescales that are harder for humans to observe. Thus ENSO and its diverse accompanying phenomena provide us with short-term, but dramatic, illustrations of the ocean's effect on climate.

22.5.3 Prolonged global warming and ocean circulation shutdown

It is natural to speculate what the long-term impacts of global warming might be on the oceanic–atmospheric system. As the Younger Dryas episode suggests, the complex and chaotic system that we call climate could flip-flop between states if perturbed strongly enough. It has been proposed that prolonged global warming, in inducing extensive melting of polar ice sheets, might flood the North Atlantic with buoyant freshwater, preventing the release of heat and the sinking of North Atlantic warm waters, and shut down the circulation pattern that extends all the way down to Antarctica.

In such a circumstance, the climate in Europe could become substantially colder, and other climate perturbations might be expected worldwide. It has even been suggested that the increased rainfall in the North Atlantic predicted by some AOGCMs could inhibit or shut down this circulation. Once shut down, it might takes tens of thousands of years for a gradual rearrangement of the ocean salinity and heat to re-establish a circulation, and there is no guarantee it would be the same as today.

What would be the result of such a shutdown? Colder temperatures along the North Atlantic would encourage the build up of ice sheets over a larger area for a longer fraction of a year. It is tempting to suggest that such a process could push the climate system in the other direction, that is, toward an ice age. However, the interrupted circulation of the ocean might inhibit the absorption of the large, human-induced carbon dioxide abundance in the atmosphere; it might even force release of additional CO_2 from the ocean. This would push the climate toward warmer states. Such a tug of war between the tendency toward glaciation versus even warmer overall climate likely would be accompanied by shifts in climate on decadal or annual timescales.

Our limited understanding of oceanic–atmospheric interactions makes such ideas at best speculative, and no convincing evidence has been seen in ocean circulation data that a Younger-Dryas-type mechanism is being triggered by glacial melting. But the general notion that ocean circulation could drive sudden shifts in the climate regime of Earth must be taken seriously enough to deserve further study. Too much data exists showing rapid and strong shifts in climate on the cusps of interglacial/glacial transitions, to ignore the possibilities.

22.6 Global warming: a long-term view

Although society is naturally most concerned about possible effects over the next 50 years (one to two generations), the longer term legacy of fossil-fuel burning is also of interest. Once released, CO_2 produced from fossil fuels will remain in the oceanic–atmospheric system for many centuries or longer, until it is sequestered on the ocean floor by biological processes and then subducted. The fossil fuel "bank" from which we make our withdrawals of carbon does not readily accept deposits. Hence, even if we significantly slow fossil-fuel burning worldwide, elevated CO_2 levels will persist over many human lifetimes, and a return to preindustrial values will not occur, by some estimates, for many thousands of years.

By the middle of the next millennium, that is, around AD 2500, atmospheric carbon dioxide levels could be as high as 1,000 to 2,000 parts per million, compared to the 389 parts per million of today. Simplified models that ignore the effects of oceans yield a climate in that time frame as warm as the Cretaceous climate described in Chapter 19. Of course, the oceans are likely to strongly modify the warming, as described in the preceding section, through absorption of carbon dioxide and transport of heat. Regardless, however, of the effect of oceans, humans will still have made a lasting impact on the climate system by introducing into the oceanic–atmospheric system, on geologically short timescales, up to 10 times the preindustrial amount of carbon dioxide. Other than episodes of high volcanic activity, and large impacts, this change to our atmosphere has no precedent in the Phanerozoic. Indeed, Nobel Prize-winning chemist Paul Cruzan has proposed that the time from the twentieth century onward be labeled the "anthropocene", when human activities began to dominate the radiative balance of the atmosphere. Like the photosynthesizing oxygen producers of the Proterozoic, we are creating something of an atmospheric revolution – and like those earlier revolutionaries, we will have to adapt to our changes.

22.7 Postscript: human effects on the upper atmosphere – ozone depletion

Another human-induced perturbation of the atmosphere involves the thinning of the layer of ozone high in Earth's atmosphere (10 to 35 km above the surface). This layer absorbs ultraviolet photons from the Sun, which are lethal to plant and animal life near the surface. The natural process of ozone destruction and formation involves the break up of ozone (O_3) molecules by sunlight to form atomic and molecular oxygen (O and O_2, respectively), and reformation of O_3 in a chemical chain involving oxygen, water, and sunlight. Certain reactive elements such as chlorine can accelerate the break up of ozone by acting as a catalyst: a single atom of chlorine (Cl) reacting with ozone produces ClO and stable O_2. The ClO is also reactive and quickly combines with another oxygen atom to produce atomic chlorine and an O_2 molecule. The chlorine atom is now free to attack another ozone molecule.

Eventually the chlorine may react with other molecules in the atmosphere to produce more stable compounds such as HCl or $ClNO_3$ that are removed from the stratosphere – but only after a large number of cycles in which the chlorine has destroyed ozone. Human-produced refrigerants called chlorofluorocarbons, or CFCs, contain chlorine. These compounds drift up into the stratosphere, where the chlorine is released by solar ultraviolet radiation. Since the mid-twentieth century this source of chlorine has exceeded natural sources such as sea salts, which is why CFCs are accelerating the loss of ozone from the stratosphere.

Monitoring by satellites and aircraft now provides a worldwide picture of stratospheric ozone variations. Since the mid-1970s an ozone hole has appeared over Antarctica, indicating enhanced destruction of ozone there. Apparently the cold winter temperatures and sluggish atmospheric circulations above the southern polar region encourage the formation of water-ice clouds in the stratosphere. The ice particles act as chemical sites where the HCl and $ClNO_3$ react with each other to form Cl_2. As spring approaches and sunlight hits the Antarctic atmosphere again, the Cl_2 is broken apart into atomic chlorine, Cl. Thus, instead of being removed from the atmosphere, the chlorine is recycled into its active catalytic form again, enhancing the destruction of ozone and producing the deep springtime ozone hole around the south pole.

The particular climate conditions over the Antarctic conspire to make that region more sensitive than other parts of the globe to the introduction of chlorine into the stratosphere. However, this does not mean the rest of the world's ozone is immune to the effects of enhanced chlorine. Ozone depletion is seen at all latitudes in the northern and southern hemispheres. Because the residence time of active chlorine in the stratosphere is decades, enhanced destruction of ozone worldwide will continue even though international agreements forced the phasing-out of CFCs beginning in the late 1980s.

Summary

Excellent data from ice cores and measurement of atmospheric CO_2 since the middle of the twentieth century shows that this greenhouse gas has increased by 40% since the time prior to beginning of the Industrial Revolution; most of this increase is the result of human activities. Other greenhouse gases have increased as well. The record of temperature, averaged over the year and over the surface of the Earth, is more difficult to interpret because of the urban heat island effect and the variable quality of instruments over time, but the global average temperature seems to have increased about 1 degree Celsius since the nineteenth century, and the increase became steeper after 1970. The results are consistent with what is expected in a simple, physically sound model of the balance between infrared and visible radiation in a cloud-free atmosphere – that is, the Earth's lower atmosphere is warming up due to increase in greenhouse gases. Complications in the exact relationship between the increase in greenhouse gases and temperature include the role of water as vapor and clouds, atmospheric convection, and the role of the oceans in absorbing and releasing heat and greenhouse gases. Detailed predictions of the role of the oceans and of changes to regional weather patterns associated with global warming are obtained though general circulation models of the Earth's atmosphere and oceans. Such models do well in predicting certain effects, like increased overall precipitation, but other aspects of changing climate are more uncertain. Crucial to the behavior of Earth's climate in the next century is the role of ocean circulation, for example, potential changes to the frequency and intensity of the El Niño phenomenon, and more speculative possibilities such as reduction in the thermohaline circulation of the North Atlantic as accelerated glacial melting and added rainfall change the surface saline content of the ocean. As we grapple with global warming, another human-caused impact on the environment, stratospheric ozone depletion, appears to have been resolved by international protocols on prohibition of industrial chemicals known to accelerate destruction of this molecule essential for reducing the amount of solar ultraviolet radiation reaching Earth's surface.

Questions

1. Consider how you, as a policy maker, would weigh the economic consequences of various responses to global warming predictions. Would you take aggressive action now or a wait-and-see attitude?
2. How would you seek to eliminate the urban heat island effect from data sets used to construct global average temperature over time from the mid-nineteenth century to the present?
3. What might the response of climate be to the oceans if, hypothetically, they extended no more than 100 meters deep (as opposed to the actual, deep-ocean situation on Earth)?
4. Do a literature search to collect evidence for shrinkage of ice over the last century from glaciers, mountain tops and the Earth's polar regions. What evidence is there that this shrinkage is not part of a normal cyclical waxing and waning?

General reading

National Research Council. 2010. *Advancing the Science of Climate Change*. National Academies Press, Washington DC.

Peixoto, J. P. and Oort, A. H. 1992. *Physics of Climate*. AIP Press, New York.

References

Bright, C. 1997. Tracking the ecology of climate change. In *State of the World 1997* (L. R. Brown, C. Flavin, H. F. French, and L. Starke, eds). W. W. Norton, New York, p. 22.

Broecker, W. S. 1995. Chaotic climate. *Scientific American*, **273**(5), 62–8.

Broecker, W. S. and Denton, G. H. 1990. What drives glacial cycles? *Scientific American* **262**(1), 49–56.

Crowly, T. J., and Kim, K.-Y. 1995. Comparison of longterm greenhouse projections with the geologic record. *Geophysical Research Letters* **22**, 933–6.

Davis, B. A. S. Brewer, S. Stevenson, A. C., and Guiot, J. 2003. The temperature of Europe during the Holocene reconstructed from pollen data. *Quaternary Science Reviews* **22**. doi:10.1016/S0277-3791(03)00173-2.

Dessler, A. E. 2010. A determination of the cloud feedback from climate variations over the past decade. *Science* **330**, 1523–7.

Enfield, D. B. 1989. El Niño, past and present. *Reviews of Geophysics* **27**, 159–87.

Houghton, J. T., Ding, Y., Griggs, D. J. *et al*. (eds). 2001. *Climate Change 2001: The Scientific Basis*. Cambridge University Press, Cambridge, UK.

Knutti, R. 2008. Should we believe model predictions of future climate change? *Philosophical Transactions of the Royal Society, Series A* **366**, 4647–64.

Manabe, S. and Stouffer, R. J. 1993. Century-scale effects of increased atmospheric CO_2 on the ocean–atmosphere system. *Nature* **364**, 215–18.

Mann, M. E., Park, J., and Bradley, R. S. 1995. Global interdecadal and century-scale oscillations during the past five centuries. *Nature* **378**, 266–70.

Meyers, S. D. and O'Brien, J. J. 1995. Pacific ocean influences atmospheric carbon dioxide. *EOS* **76**, 533.

Mitchell, J. F. B. 1989. The "greenhouse" effect and climate change. *Reviews of Geophysics* **27**, 115–39.

Mortensen, L. L. (ed.). 1996. *NOAA Global Change Education Resource Guide*. US Dept. of Agriculture, Washington, DC.

Oerlemans, J. 1994. Quantifying global warming from the retreat of glaciers. *Science* **264**, 243–5.

Stone, P. H. and Risby, J. S. 1990. On the limitations of general circulation models. Center for Global Change Science, MIT, Report 2, unpublished.

Subcommittee on Global Change Research. 1995. Forum on global change modeling. *U.S. Global Change Research Program*, USGCRP Report 95–02.

Toggweiler, J. R. 1994. The ocean's overturning circulation. *Physics Today* **47**(11), 45–50.

Trenberth, K. E., Houghton, J. T., and Meira Filho, L. G. 1996. The climate system: an overview. In *Climate Change 1995: The Science of Climate Change* (J. T. Houghton, L. G. Meira Filho, B. A. Callander, N. Harris, A. Kattenberg, and K. Maskell, eds). Cambridge University Press, Cambridge, UK, pp. 51–65.

Wahlen, M., Allen, D., Deck, B., and Herchenroder, A. 1991. Initial measurements of CO_2 concentrations (1530 to 1940 AD) in air occluded in the GISP ice core from central Greenland. *Geophysical Research Letters* **18**, 1457–60.

World Meteorological Organization. 2011a. *The status of the global climate in 2010*. WMO no. 1074, Geneva, Switzerland.

World Meteorological Organization. 2011b. *Greenhouse gas bulletin*. No. 7 WMO Geneva, Switzerland.

23

Limited resources: the human dilemma

Security is mostly a superstition. It does not exist in nature, nor do the children of men as a whole experience it.

HELEN KELLER

Introduction

Only in the last century has humanity's command of technology and energy made it possible to feel a sense of security unknown for most of human history. And yet, ironically, this new sense that we can obtain and control what we need to make life lengthy and of high quality comes just as we find ourselves depleting what were once thought to be virtually inexhaustible resources. As our numbers grow, how will humankind continue to sustain the industrialized civilization that has made high standards of living – or the aspiration for such standards – a staple theme of the last century? This chapter addresses briefly the issues of future supplies of food, energy, and material resources.

23.1 The expanding human population

Population growth is the root cause of human-induced global warming and depletion of non-renewable resources. From the beginning of humankind to just over 100 years ago, the world's human population was less than one billion. Our planet now holds nearly 7 billion persons thanks to medical advances and growth in agricultural productivity throughout the twentieth century. The growth rate will take us to just over 9 billion by the middle of this century. The net increase in population over the last decade amounts to about 150 million people a year. Growing population is a two-edged sword. Increasing numbers of people, supported in adequate living standards by advancing technology, represent an expanding reservoir of personalities, innovative ideas, and the creative seedcorn for future developments in both technological and humanistic spheres of existence. On the other hand, unbridled population growth that outpaces technological developments designed to stem its negative impacts could push humanity into a downward spiral of resource depletion, decreased overall living standards, and ever more profound alteration of natural systems by human activities.

Approximately 30 countries, most of Europe along with Japan, have achieved a roughly zero population growth rate (actually, an annual growth rate of less than 0.3%, as defined by the Washington, DC-based Worldwatch Institute). Many of these countries did so without a deliberate effort: the fall in birth rate reflects rising living standards and increasing career opportunities for women. In other cases, populations are declining in the context of failing health and living standards. Russia is the principal example, with a negative annual growth rate of 0.5%. In total, the world population growth rate is declining (Figure 23.1), overall a positive sign that we may be entering a period where technologies for growing food and providing energy can catch up with the prodigious population growth of the twentieth century.

The benefits to a given country of a stable population can include the ability to increase exports of consumables, which otherwise would always be "catching up" with the increasing demands within the country itself; potential export of European grain is one example. Not all prosperous nations have achieved near-zero population growth: that of the United States continues

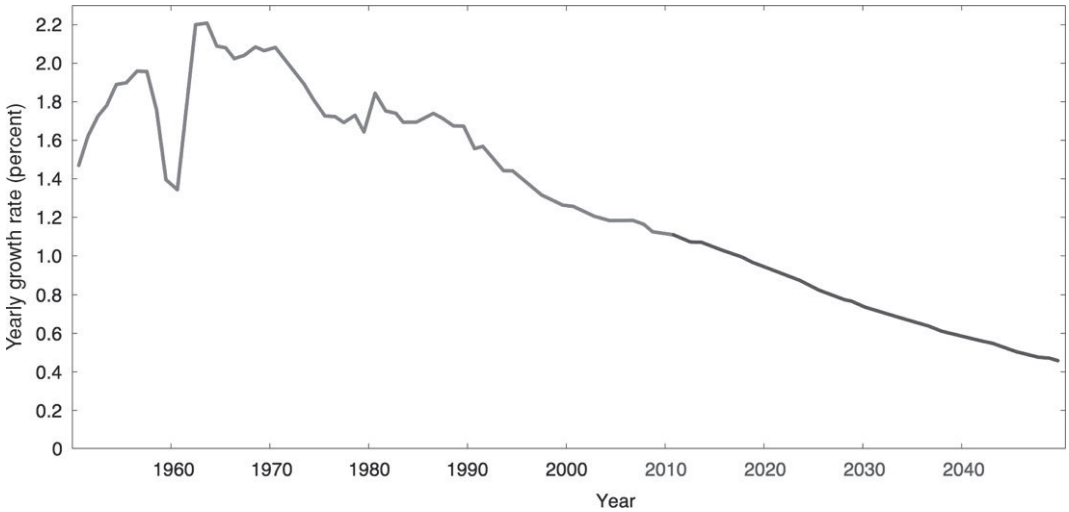


Figure 23.1 Yearly growth rate of the world's population in percent. Numbers after 2010 represent a projection. Figure based on data from the US Census Bureau and constructed by Securiger.

to rise at a rate of 1% per year (twice that of China), and there is as yet no economic incentive to trim the growth rate. In most developing countries, the reduction in birth rate is tied to the economic status of women. Many traditional societal structures leave women bereft of decision-making or economic power and force them into the role of bearing and raising large numbers of children. So-called "microenterprise" programs provide seed loans for women of developing countries, enabling them to begin their own businesses. In at least some cases the result has been to break the vicious cycle of poverty and large families.

In the remainder of the chapter we focus on issues connected to population growth that illustrate the dilemmas humankind faces as its numbers continue to expand: food, energy, and material production. These are intimately coupled to the question of human-induced global warming, because consumption of fossil fuels may drive global climate change, and food production is sensitive to the net results. At the start of the twenty-first century, for the first time in history, our problems of adequate food and energy supply are global in nature, with no prospect that a long-term solution lies in finding "other places on Earth" where new land and new resources are available.

23.2 Prospects for agriculture

The invention of agriculture occurred sometime in the late Pleistocene or early Holocene, and undoubtedly in several different places at various times. Major gains in agricultural productivity have occurred at various times throughout human history. Cro-Magnon in the Europe of 30,000 years ago may have required 1,000 hectares (or 10 square kilometers) of land to support a single family, largely by hunting. By the early Holocene, agricultural techniques using domestic animals brought this down a factor of 100, to 10 hectares per family. Medieval European farming practices supported one family on a hectare of land. Nineteenth century Asian rice-growing techniques reduced this by another factor of 5. Among today's most intensive agricultural systems, Japanese rice farmers support a family on 0.1 hectare of land – 10,000 times less area than the Pleistocene humans required.

The Industrial Revolution fundamentally changed the techniques of agriculture beginning in the mid-eighteenth century in England. The development of engines to harness coal and eventually oil created a transportation infrastructure that accelerated movement of goods, including agricultural commodities and tools, between urban and rural areas. It also enabled the mechanization of agriculture itself, improving yields. The most recent agricultural innovation is the "Green Revolution," beginning in the mid-1960s and characterized by development of novel high-yield varieties of dwarf rice and wheat, along with expanded use of irrigation, chemical pesticides, and fertilizers. The resulting increase in yield from that time onward (Figure 23.2) has helped to sustain a growing population without an increase in the percentage of the population facing chronic hunger; that percentage has shrunk to about 15% compared with over 30% in 1969. However, this means that 1 billion people in the world chronically do not get enough to eat.

The increasing yield was aided for some time by a continuing expansion in the area of land on the planet brought under agricultural cultivation. This increase has now ceased in Eurasia and the Americas, and in Africa further dramatic increases in farmland are unlikely. In much of the world, intensive agricultural practices have led to severe soil erosion and loss of topsoil, so that the ability to retain today's cultivated farmland is in serious doubt. In developing countries, loss of agricultural land to industrialization is significant. This is particularly evident in China, where 100 million rural workers migrated during the 1990s into cities hoping for better jobs and a higher living standard. A conservative estimate of the amount of farmland paved over to accommodate these new urban dwellers is 435,000 hectares: sufficient to feed millions of people. In wealthy countries such as the United States, a different social phenomenon is destroying arable land: significant amounts of the best farmland are being lost to development of new suburban communities populated by a middle-class trying to escape the problems of urban environments.

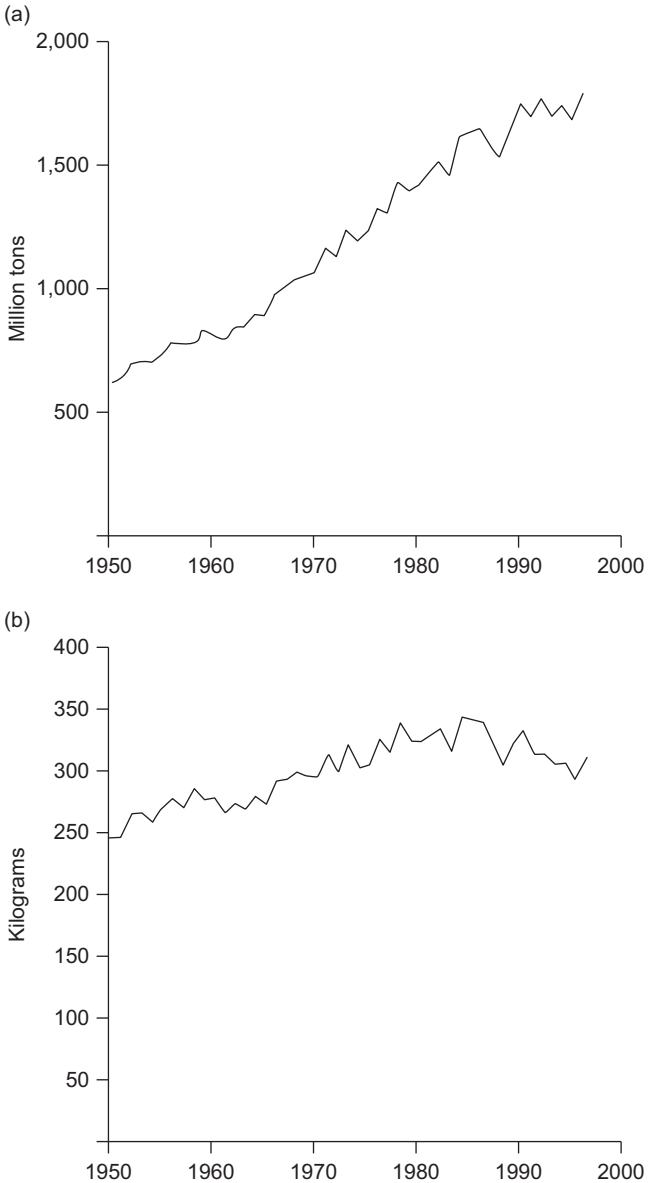


Figure 23.2 Estimates of total worldwide grain production (a), and grain yield per capita (b) from 1950 to 2000. Data source: World Resources Institute (http://www.wri.org/).

Current food production is delicately balanced with increasing population pressure and desire for higher living standards worldwide. Large increases in food production come with a price: increased energy consumption and increased release of both carbon dioxide (from energy consumption associated with mechanized farming and transport of agricultural hardware and products) and methane (from agricultural activity), which may intensify global warming. Progressive soil erosion due to intensive farming practices reduces arable land. Perhaps most ominous, worldwide agricultural production potentially is among the most sensitive of human endeavors to accelerated global warming. As discussed in Chapter 22, a general prediction of global circulation models is increased soil dryness, particularly in equatorial regions, due to rising temperatures and an inability of increased rainfall to compensate for the greater amount of evaporation. Consequent loss of productivity of agricultural lands could accelerate the trend of decreasing harvestable area per person, and blunt the positive impacts of new agricultural technologies, such as development of new crops by genetic engineering. Although some have argued that higher latitude regions could be developed into farmland in a warmer global climate, much of this land contains very thin, poor soils, the yield of which is likely to be limited.

It is difficult to predict the potential for net loss of agricultural land over the next half century because of the sensitivity to details of the climate models, but the trend is clear. The next agricultural revolution of necessity will be characterized by techniques designed to protect the current world inventory of agricultural land against loss both by destructive farming techniques and by human-driven or human-accelerated global climate change.

Earth's surface area is finite, and humanity's ability to feed itself in the face of its ever-increasing numbers depends on driving up the food yield per acre. In some respects, our entire post-Pleistocene history has been characterized by innovations that have made increasing yields possible. Although we are nowhere near any fundamental limit on producing food from the overall inventory of surface organic material, neither are we guaranteed a future of increasing abundance. Human ingenuity may or may not find new ways to squeeze even more food from the Earth. In a general sense, our ability to do so may depend on utilizing energy sources that have less impact on the global climate than those we now employ. It is to the issue of energy sources and their limitations to which we now turn.

23.3 Energy resources

23.3.1 Fossil fuels

Coal and oil, along with natural gas, have been the traditional primary sources of energy since the nineteenth century. These *fossil fuels* are derived from long-dead organisms whose content of carbon and hydrogen is buried in the earth. Coal is largely derived from the decay of organic matter from forests of the Carboniferous period. Large quantities of dead plant matter collected in swamps, undergoing slow decomposition. The pressure of accumulated layers of sediments over geologic time increased pressures and temperatures in the organic layers, forcing out moisture and other materials. What was left were beds of compact carbon-rich material, *coal*, interleaved with sandstones, shales, or other sediments. The highest temperatures and pressures produced the highest grades of coal, such as anthracite, which is nearly pure carbon. Lower heating and pressurization produced bitumous coal, an industrial fuel with less carbon per gram than anthracite.

The development of coal as an energy source was pioneered in Great Britain. By the end of the nineteenth century, scarcity in wood supply coupled with the design of increasingly efficient steam engines drove the large-scale production of coal. British coal production was two-thirds of the world's production capacity of this fuel through the middle of the nineteenth

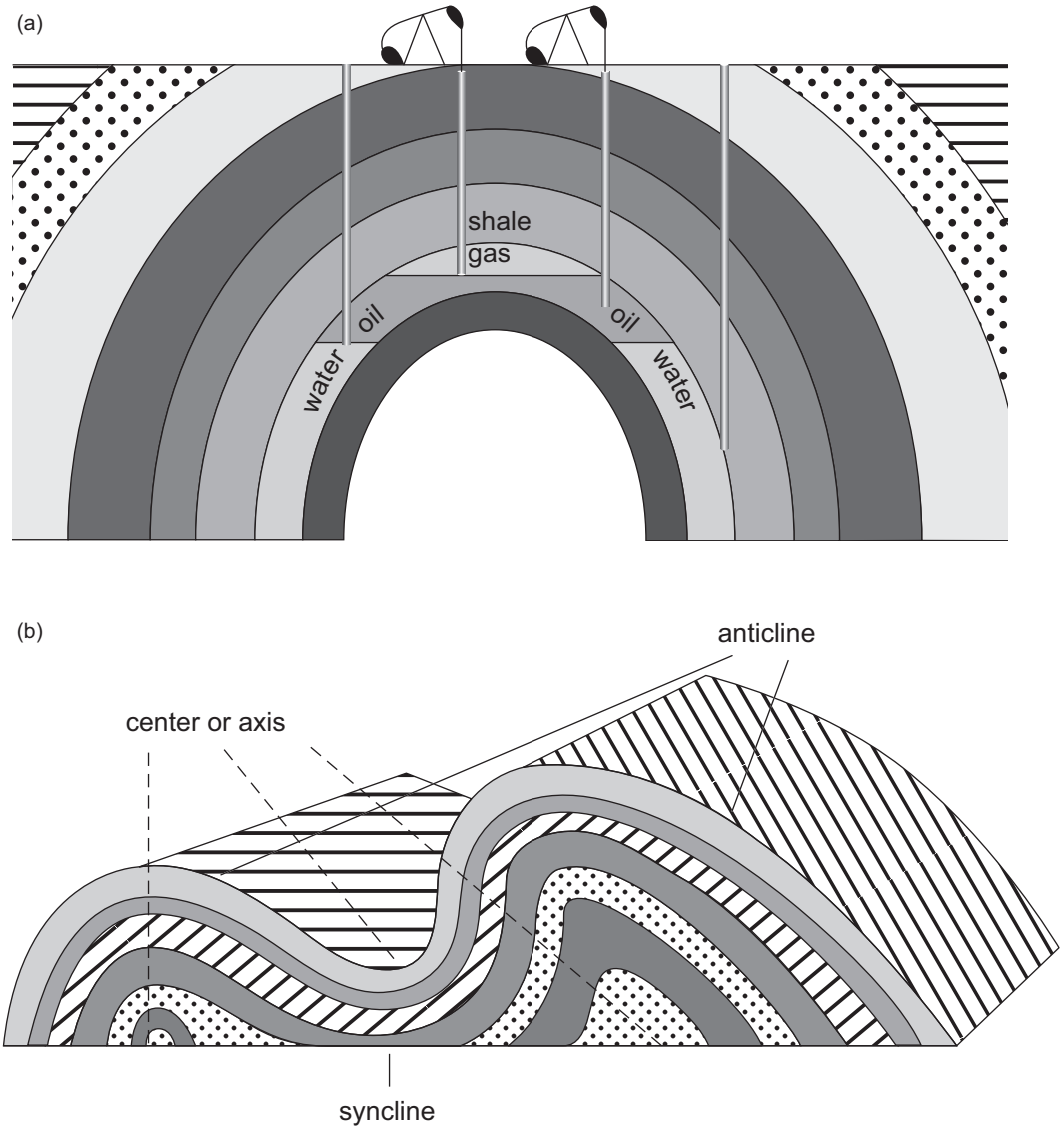


Figure 23.3 (a) Cross section through sedimentary layers showing reservoirs of oil and gas. Here the hydrocarbons have been transported by water from shales into porous sandstones. The oil floats above the water and is trapped. (b) Anticlines – convex folds in sedimentary layers – are attractive regions to explore for fossil fuels, because the low-density oil migrates upward toward the top and collects below an impermeable layer.

century, eventually to be succeeded by the United States. Coal is a relatively "dirty" fuel, producing large amounts of emission per unit of energy derived. Problems with air pollution from coal occurred in England as early as the thirteenth century, when it was burned directly in homes as a source of heating and cooking; Queen Elizabeth I at the time banned its use in brewhouses within a mile of the court.

Oil production began in the United States in the mid-nineteenth century, its first use being limited to oil lamps and a few other purposes. However, the invention of the automobile spurred the increased production and refinement of oil, to be followed by natural gas as a somewhat cleaner substitute for many applications. The widespread use of oil and gas transformed the worldwide economic landscape, creating rich empires out of formerly impoverished countries that happened to sit atop major oil reserves. The formations of oil and natural gas are somewhat more complicated than for coal, and are illustrative of the natural processes associated with formation of energy reserves (Figure 23.3):

1. Living organic matter is produced at the surface by biological processes, beginning with photosynthesis.
2. The organisms die and the organic matter must be buried in sediments before it is destroyed by reaction with oxygen at the surface.
3. Parts of the organic material undergo reactions to form petroleum and/or natural gas (methane and other volatile carbon-hydrogen compounds).
4. The oil and gas migrate into permeable (porous or cracked) beds of rock.
5. The beds of rock are folded or faulted in such a way as to trap the oil and gas.

6. The above must happen recently enough so that the traps have not been heated by tectonic activity, which would effectively burn the oil and gas into a nonusable form. Structural deformation also must be limited, because this could cause cracks that would allow the oil or gas to escape. Thus most oil and gas geologically is relatively young (post-Cretaceous).
7. The beds must remain buried until the present day because exposure dissipates the oil.
8. Recoverable oil comes from beds that retain a high permeability. Clays or other mineral cements, if they get into the pores before recovery, prevent the oil from being pumped out.

23.3.2 The challenges of fossil fuels

Fundamentally, fossil fuels represent stored solar energy. Fossil fuels have their origin in plants and animals that lived on Earth in the past. These organisms, of course, were born and thrived as part of a food chain whose base is the photosynthetic production of sugars by plants, using sunlight as energy, and primarily carbon dioxide, water, and nitrogen as raw materials. The vast majority of Earth's crustal carbon is stored as limestone (carbonates) and in shales, so that fossil fuels represent a very tiny fraction of the buried crustal organic matter.

The use of fossil fuels by human beings represents an acceleration of the portion of the carbon cycle that returns buried carbon to the atmosphere as carbon dioxide (Chapter 14). Chemically, we are extracting reduced carbon (hydrogen-rich carbon, or hydrocarbons) from the crust, oxidizing it (combining it with oxygen) by burning it to yield energy, and then depositing the oxidized carbon (carbon dioxide) into the atmosphere. We are, in effect, greatly speeding up the portion of the carbon cycle wherein carbon trapped in sediments is subducted to high temperatures, converted to carbon dioxide, and emitted from volcanoes. Our contribution is a dominant one: in a matter of decades we will double the carbon dioxide content of the atmosphere. How long we can do so depends on the supply available, which is highly uncertain. The proven reserves of oil as of 2011 are shown in Figure 23.4, but this is thought to be a small fraction of what remains in the crust and is recoverable. Estimates of roughly a century or so to deplete the world's accessible oil reserves, at current usage rates, might be extended by a factor of several depending on the amount of undiscovered petroleum and use of other fossil fuel resources. Coal is much more abundant, but also much more polluting in terms of carbon dioxide produced per unit energy generated.

In the past decade, natural gas (methane with an admixture of primarily ethane, C_2H_6) has become increasingly important. Most natural gas has two sources – buried organic material that is transforming to coal and oil (thermogenic) and living microbes that produce methane in shallow land and sea sediments, bogs, and even landfills (biogenic). Vast fields of gas in shales, particularly in North America, may change the geopolitical balance of suppliers versus consumers of fossil fuels. However, extraction of natural gas, particularly involving injection of water and chemicals under high pressure to fracture rock and release the gas in a process called hydraulic fracturing or "fracking," may inadvertently contaminate groundwater drinking supplies.

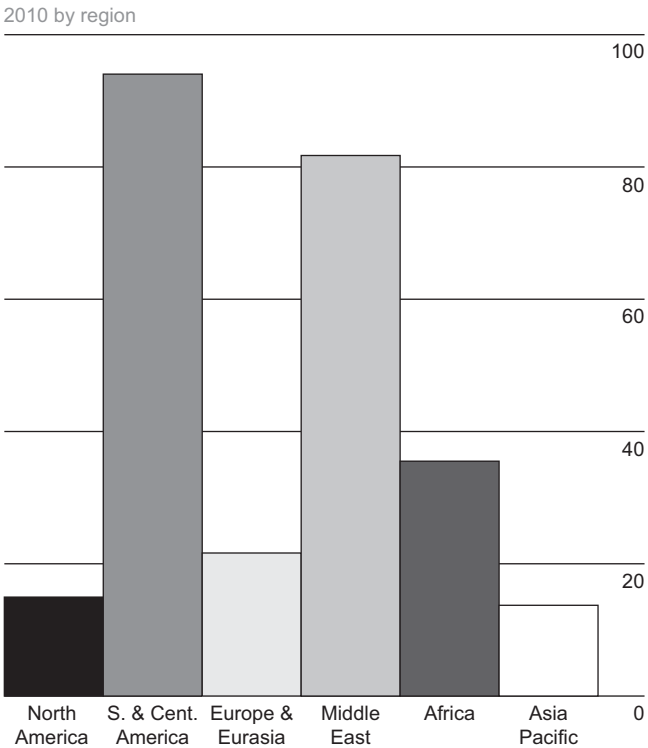


Figure 23.4 World proven oil reserves by region normalized by production. Numbers on the right are the number of years for which the proven oil reserves will meet production. This is not the number of years to the depletion of all recoverable oil, since there remain new reserves to be discovered or unproved ones in process of becoming proven (that is demonstrated with high confidence to be recoverable). Source: BP Statistical Review of World Energy, June 2011. www.bp.com/statisticalreview.

Some of the biogenic natural gas is trapped as gas hydrates, or clathrate hydrates. These are compounds of water ice stabilized by the trapping of other molecules, such as methane, in void spaces in the ice. If sufficient methane (produced, for example, biologically by microorganisms in seafloor sediments) is in contact with water at the pressures found near the ocean floor, the water will freeze even though the temperature is above the usual freezing point of 0°C. The presence of methane induces the freezing, the methane becomes trapped in the ice, and this gas-suffused ice is stable in the sediments of the seafloor until extracted by drilling. There is an intriguing possibility that large amounts of methane gas might be locked in seafloor sediments and permafrost regions in this way. How much methane exists to be tapped as a natural gas energy source is controversial. Also controversial is the issue of safety – some extraction techniques could trigger large landslides in seafloor sediments, releasing large quantities of methane suddenly and exacerbating global warming. Even on a small scale, deliberate or accidental extraction during deep sea drilling can lead to disasters, such as that of the British Petroleum *Deepwater Horizon* explosion in 2010, caused by inadvertent release during drilling of methane from a pocket of what was likely gas hydrate. The explosions killed 11 workers and caused 3 million gallons of crude oil to spill into the Gulf of Mexico, the worst oil spill in US history.

This human contribution to the carbon cycle is a novel one that began only in the last few centuries. Prior to the Industrial Revolution, the use of wood as a fuel predominated. Because wood is a product of biochemical processes in living plants that directly extract carbon dioxide from the atmosphere, burning of wood adds no net carbon dioxide to the atmosphere as long as the replenishment of trees balances their consumption. In fact, this balance probably has not been achieved for many centuries, the deforestation of the eastern United States beginning in the eighteenth century being a prime example of net consumption of wood. The post-Industrial Revolution world simply could not rely on wood as a principal energy source, because demand would enormously outstrip supply. Thus, we are committed to the use of fossil fuels, and hence to steady increase in the amount of carbon dioxide introduced to the oceanic–atmospheric system, or we must consider sources of energy that do not rely on the oxidation of reduced carbon extracted from the Earth's crust. Such *alternative energy sources* have their own pitfalls and limitations.

23.3.3 Alternative energy sources

Biofuels are fuels generated by biological carbon fixation – what happens when growing crops. And, indeed, biofules include foods like corn or sugarcane. Biofuels can be used in the form of gasoline, diesel, and airplane fuel, among others. Because they are made by plants extracting carbon dioxide from the atmosphere, in principle they do not contribute directly to increased levels of CO_2 in the atmosphere, although indirectly they do because of the energy cost of farming, refining, and transport. A criticism of biofuels is that they use crops and cropland that might otherwise have gone to food production, although using these fuels to blunt the climatic effects of fossil fuel use might in the long run help food production.

Solar energy involves materials that absorb sunlight and store the energy through chemical reactions. Solar energy is nonpolluting, but the amount of energy generated depends on the area of the solar panels. To make solar energy the primary contributor to electrical energy in North America could require thousands of square kilometers of solar farms. One could imagine the deserts of the western United States arrayed in this way, but the initial infrastructure investment is too costly at present. Further, the idea of covering vast natural areas with solar arrays probably would meet with public disapproval. Likewise, space-based arrays would be very expensive, and the beaming of energy to Earth, by microwaves, could cause environmental or health hazards. Individual use of solar arrays by homeowners to reduce their dependence on community electric power grids is popular in dry, sunny locations, and may become more so as technological improvements drive down the cost of such systems and increase their efficiency.

Geothermal energy can, in principle, be relatively nonpolluting, but the known geothermal reserves that can be economically tapped are not sufficient to supply current world energy needs. *Hydroelectric power* also falls short by a wide margin of supplying world energy needs.

Nuclear fission, the splitting of uranium to generate heat, does not contribute to carbon dioxide production and hence is not a contributor to global warming. Also, the supplies of uranium oxide are such as to provide more energy than the world's coal reserves. However, nuclear reactor accidents such as Fukushima Daiichi (2011), Chernobyl (1986), and Three Mile Island (1979) have led to strong public sentiment against the proliferation of this type of energy generation process. Thus, in spite of continuing improvements in designs, public acceptance of fission as an energy source remains relatively low. There is also the unresolved problem of disposal of waste products from such reactors. Nonetheless, a number of countries successfully and safely rely on nuclear fission as a primary energy source.

Fusion reactors mimic the Sun's energy-generating nuclear fusion mechanisms and leave behind much less radioactive waste per unit of energy produced than does fission. Hydrogen can be fused to form helium, but it is easier to fuse deuterium. Deuterium is available from seawater in huge quantities and would provide a supply of clean energy to satisfy the world for centuries to come. These reactions require enormous temperatures, however, and sustained generation of energy by fusion is beyond the forefront of today's technology. Fusion technology in the United States and Russia is perhaps 100 years away from generating energy cheaply enough to drive industry in that direction. In fact, current experiments still require more energy to run than they make during the fusion reaction, and fundamental physical problems associated with efficiently extracting energy from the fusion reaction (while sustaining it) may never be solved in the view of some experts.

23.3.4 Energy use in the future

Potential changes in the use of different types of energy sources predicted in 1986 is shown in Figure 23.5. Despite the vintage nature of the graph, the prediction of the relative importance of the different sources in the coming decades has not changed much. We are still, and will be for the foreseeable future, a fossil-fuel based society. Only biofuels have the possibility to change this picture, but not in a fundamental way. Coal offers, by far, the largest amount of usable energy resource, but extraction involves environmental damage (strip mining) and health hazards to workers. Also, coal tends to produce at least twice the amount of carbon dioxide per unit of energy, when burned, as does oil. Natural gas (primarily methane) produces the least amount of oxidized carbon per unit of energy generated.

Demand for and burning of fossil fuels will continue for many decades even under the assumption of stringent conservation measures. Present growth in the demand for energy worldwide is roughly 2.5% per year. Based on the fraction of fuels that emit carbon, the growth in energy demand corresponds to a doubling of the present atmospheric carbon dioxide content in less than a century. Increased use of low-carbon fuels, biofuels, or alternative energy sources can decrease the rate, but the prospect for dramatic reduction in carbon emission will probably require stringent efforts in conservation.

23.4 Economically important minerals

In addition to energy, our civilization requires a suite of metals and other elements as the raw materials for products and processes. Mining and processing of metals extends back to earliest

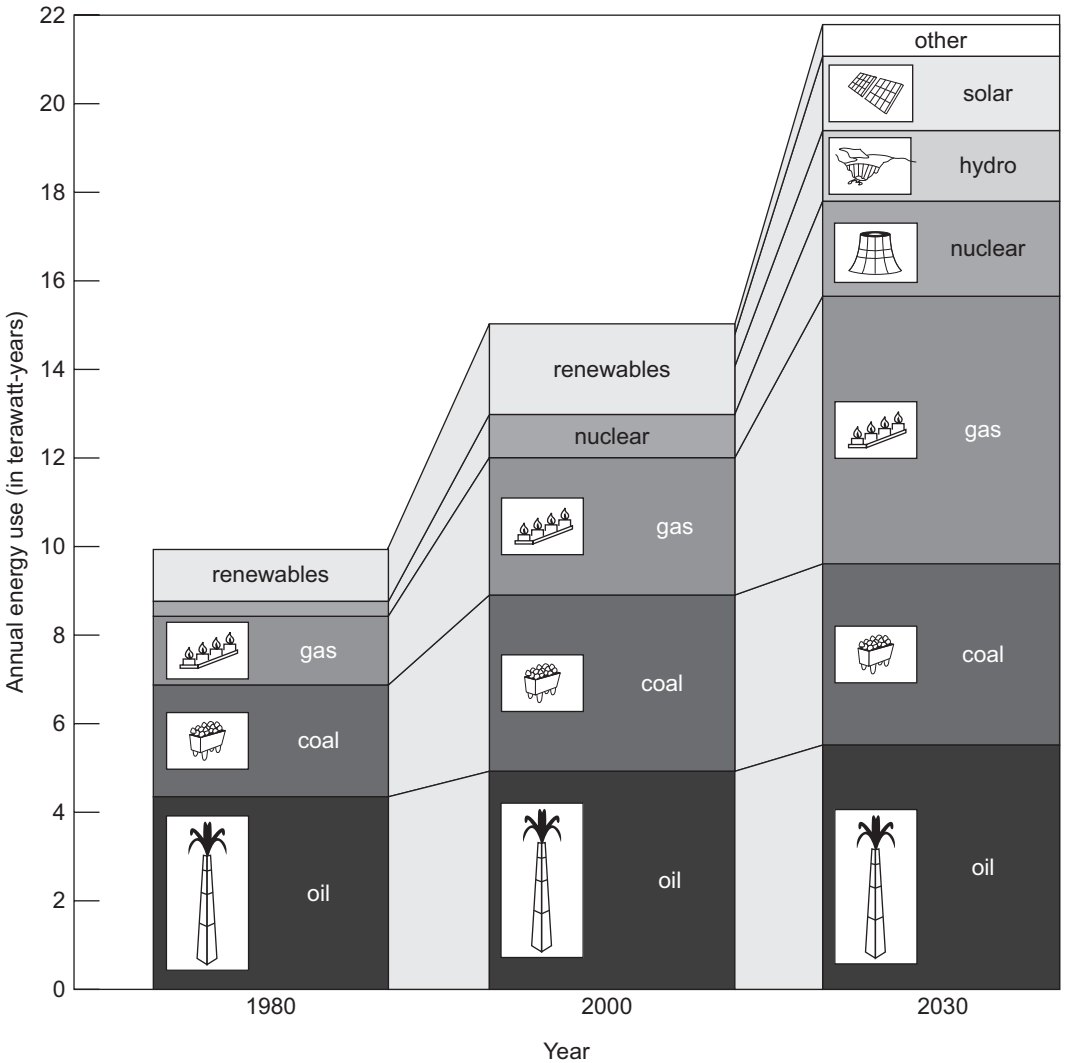


Figure 23.5 A 1986 projection of the change in world dependence on various sources of energy from the year 1980 through 2030. Renewable energy is broken out into solar, hydroelectric, and other sources for the year 2030. Annual energy use is given in units of terawatt-years, which is the energy expended in generating a trillion watts of power for a year, or roughly the consumption of a billion tons of coal. Despite the age of the graph, the prediction of the relative importance of the different sources in the coming decades has not changed much. From Rogner (1986).

organized civilizations. The record of such activities is preserved in ice sheets and peat bogs. Cores extracted from, for example, the Greenland ice sheet have been analyzed for the presence of copper, lead, zinc, etc., using sensitive laboratory techniques. The origin of these metals in high-latitude ice sheets lies in the efficient transport of pollutants through the atmosphere from low-latitude sites (where the highest concentration of humans have existed) to other latitudes. Atmospheric circulation models suggest that the Greenland deposits of metals in ice provide a good record, as a function of time, of the amount of these metals introduced into the atmosphere.

The records of metal pollution in the atmosphere compare favorably with the timing of civilizations that were heavy users of metals, such as the Roman Empire, the Sung dynasty of China, and the world of the post-Industrial Revolution. Lead, copper, arsenic, antimony, and other metals all begin to exceed their natural background levels beginning about 2,500 years ago. Figures 23.6 and 23.7 summarize from a suite of sources the history of lead and copper production, as examples. Interestingly, the ice-core data show a difference between the two metals in terms of modern-day versus ancient atmospheric pollution. In the case of copper, modern smelting techniques are much more efficient than ancient ones, and cumulative large-scale copper pollution of the atmosphere thus does not track the actual use of the metal. The same is not true for lead; a massive increase in atmospheric lead pollution accompanied tremendous growth in use of the automobile after World War II. This increase has been reversed in the past couple of decades with the introduction and widespread use of unleaded gaso lines.

Large-scale mining operations, essential to extracting the raw materials that are the foundation for the material goods of our civilization, carry with them environmental and health costs. Smelting copper, for example, although more efficient today

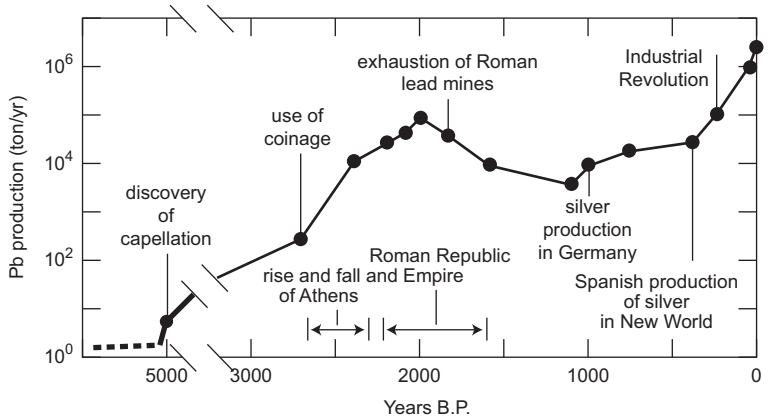


Figure 23.6 Lead production over the past 5,000 years, from analyses of Greenland ice cores. Hong *et al*. (1994) by permission of American Association for the Advancement of Science.

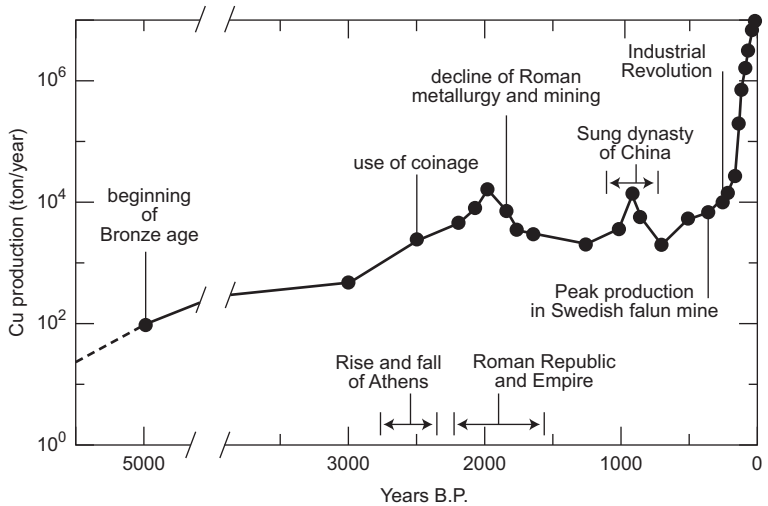


Figure 23.7 Copper production over the past 5,000 years, from analyses of Greenland ice cores. From Hong *et al*. (1996) by permission of American Association for the Advancement of Science.

than in the past, produces substantial hydrocarbon and other emissions that represent significant sources of pollution in some parts of the world. Open-pit mining, used to access deep copper sulfide deposits that cannot be extracted economically by other techniques, leaves behind a large pit and piles of excavated earth. A medium-size copper mine might involve a pit one to several square miles (two to eight square kilometers), surrounded by perhaps 5 square miles (13 square kilometers) occupied by piles of removed and processed dirt.

In the United States, an example of a mineral-rich industrialized nation, significant conflicts arise as mining companies seek to develop ore bodies that happen to be located in areas of high scenic value and heavy recreational use. Such conflicts are becoming more common and more ironic: a wealthy nation that has the resources to set aside large areas as public lands, particularly in the western United States, and the time to enjoy outdoor recreational activities in such places, finds itself faced with the necessity of extracting mineral resources from those same lands, usually with destructive consequences. Current environmental laws do not require mining companies to backfill open pits with the overburden piles, on the basis that such requirements would be highly burdensome economically. Mining in developing nations, where environmental laws are less strict, carries with it even more risks for the health of nearby residents and workers. Recycling of the current inventory of metals may reduce the demand somewhat, but ultimately new technologies must be applied to make hard-rock mining less destructive of the land, and to enable reclamation techniques that are at once more economical and more thorough. Recent progress in reclamation of areas strip-mined for coal demonstrates that technology can be applied to lessen the impact of mining activities on the land.

23.5 Pollution

All processes create waste products by virtue of their adherence to the second law of thermodynamics. A steam engine cannot extract all of the heat from the source of hot water. Organisms cannot utilize all of the organic materials they ingest as food;

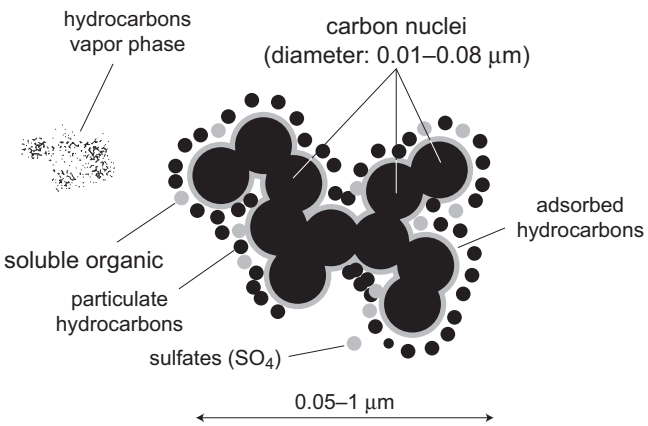


Figure 23.8 Hydrocarbon particles produced by the burning of fossil fuels. From Zmirov (1996) by permission of Les Editions des Physique.

some must be removed as waste. Likewise, all industrial processes emit some form of waste, no matter how efficient the processes may be. These wastes, emitted into groundwater or the air, represent sources of pollution that carry potential health risks.

In the case of air pollution, there are two major kinds of pollution causing health risks. Automotive and industrial emissions create nitrous oxides (NO_2), which when combined with molecular oxygen (O_2) produce NO and ozone. Ozone in the stratosphere blocks harmful solar ultraviolet radiation from reaching the lower atmosphere and surface. However, ozone produced in the lower atmosphere, and breathed by humans, can cause health problems. Ozone is a strong oxizider (donator of oxygen) that can produce very reactive *free radicals*, which directly alter biological structures. High concentrations of ozone (exceeding 300 parts per billion, or ppb) modify the structure of the mucous membranes in human respiratory tracts, particularly in the smaller, terminal airways that are crucial for transferring oxygen to the blood. In moderate concentrations (100–300 ppb), which are experienced in large cities with pollution problems such as Mexico City, inflammation of respiratory tract linings and some modification of mucus is seen. The second source of health risk is the industrial and automotive emission of fine particulates, a product of the combustion of hydrocarbon fuels, particularly diesel. These fine particulates (Figure 23.8), lodged in the lungs, aggravate existing respiratory diseases and may contribute to lung cancer.

Over the past several decades, regulatory processes have lowered emissions per unit of energy, but increasing populations and greater awareness of the link between air pollution and pulmonary disease have heightened concerns over health risks of such pollution. In Mexico City, two days of ozone levels above 110 ppb lead to an increase in asthma-related hospital visits by 68%. The costs of such increases in healthcare needs are borne by the public at large in most countries, because healthcare costs are distributed among the populace by governmental or private health insurance structures. Thus, the issue of air pollution becomes an economic one: should industry bear the costs of air pollution through requirements to implement schemes to reduce emissions, or should the general public assume the costs through pollution-related increases in healthcare needs? This difficult issue continues to be debated, but the trend is a gradually increasing imposition of requirements on industry, driven in part by public dislike of the "brown clouds" that hover over most industrial cities of the world. As energy use increases in a growing world population, political pressure will be sustained to find technologies to continue to increase efficiency and reduce pollutive waste output.

23.6 Can we go back?

Some have argued that the solution to our problems is to "go back to nature." The current 7-billion-person population is sustained against collapse by technologically based agricultural and transportation systems. To abandon technology is to regress at least back to the first agrarian societies (10,000 years ago), which were capable of sustaining less than 1% of the present human population at a much lower average standard of living than today's. The enormous and rapid increase of death rates relative to birth rates that would of necessity accompany such a regression would be regarded as a tragedy whose enormity has never before been experienced. The whole history of humankind is a story of the invention and application to the problems of living of technologies of increasing sophistication.

The very reasonable demand for a good standard of living by all the world's people will require coordination of energy, material, and technological resources among the nations. Although some economists argue that conservation measures will unduly hurt the economic engines of the industrial countries, in the long term such engines will run down unless new sources of energy are discovered and new technologies invented. Efforts now in developing countries to establish energy-efficient, high-technology production and transportation systems are mitigating the global impact of their increasing standard of living; such efforts will need to be made in the industrialized countries as well. Populations in both types of countries will need to educate themselves so as to be able to make reasoned choices with respect to new technologies, and to avoid the excessive risk aversion that too often replaces intelligent decision making.

The last half of the twentieth century was witness to two great threats to the human family: global nuclear war and overpopulation. With the former perhaps behind us, humankind's greatest challenge will be to face the latter humanely.

Summary

Population growth is the root cause of global warming and depletion of resources that may be required for future generations. Since the population will grow, and people everywhere will want to have a good standard of living, energy and resource demands will continue to increase for the foreseeable future. Food production per capita has remained steady or increased over the last few decades, but the distribution of food resources is uneven and undernourishment or starvation remains a daunting problem for much of the world's population. As global warming and urbanization progress, the ability to maintain food production volume may become more difficult. The primary energy sources since the Industrial Revolution are fossil fuels, in particular coal and oil – the processed remains of living organisms buried and subjected to heat and pressure. More recently natural gas – also a fossil fuel but produced as well by methanogenic organisms living today – has gained ascendancy. The increasing use of biofuels – made from agricultural products – takes some pressure off of fossil fuel use but consumes resources that might otherwise be used for food production. Alternatives to burning fossil files will increasingly be required if we are to slow the increase of human-generated CO_2 into the ocean–atmosphere system. Nuclear energy (fission), solar energy, wind and geothermal energy are all alternatives in use today, with various drawbacks from safety concerns (nuclear) to available sources (wind, geothermal) to available land area (solar). Nonetheless, there will continue to be an increase in the use of alternative fuels as remaining reservoirs of fossil fuels are depleted and the price of such extraction increases. Exotic forms of energy (fusion, beaming of solar energy from space) are not likely to impact the mix of energy resources for a long time to come. Mineral resources are also important for our global industrial society, and while extraction techniques have continued to improve – allowing for better reclamation of the mined land – the ecological and economic impacts of local and regional mining booms remain a challenge. Increased recycling of existing metals will alleviate this problem, but strong economic growth – such as that experienced over the last decade in China – will continue to drive up prices and make extraction from new mines attractive. All industry activities create pollution of some kind, with concomitant health problems, and so the challenge of increasing population and growth in the standard of living will be to make resource use more efficient and to maximize the capture of waste before it is introduced into the air and groundwater. Facing these challenges is not optional, since the alternative – to turn back the clock – would lead to starvation and suffering on an unprecedented scale.

Questions

1. Why do you suppose that fusion is such a challenging mechanism for generating large-scale energy for human use?
2. Develop a set of arguments in favor of a policy of encouraging continued human population expansion, under the assumption that technology will solve standard of living problems and big families provide a better chance that parents will be taken care of in their old age.
3. Look up the concept of the "tragedy of the commons" developed by ecologist Garrett Hardin in 1968, and describe some examples of the effect in your own community (or in a more distant region for which you have some experience or knowledge).
4. The stability of methane gas hydrate depends on the pressure and temperature under which it is stored. By consulting some papers on gas hydrates, describe what might happen to these deposits as the oceans warm, or alternatively, as one tries to extract the hydrates from upper sediments proceeding downward. By how much might release of all the methane in known gas hydrate reserves increase the mean radiative forcing of the atmosphere (Chapter 22).

General reading

Starke, L. and Mastney, M (eds). 2009. *State of the World: 2010*. W. W. Norton, New York.

Williams, G. R. 1996. *The Molecular Biology of Gaia*. Columbia University Press, New York.

References

Brown, L. 1997. Facing the prospect of food scarcity. In *State of the World 1997* (Brown, L. R., Flavin, C., French, H. F., and Starke, L. eds). W. W. Norton, New York.

Campbell, C. J. and Laherrère, J. H. 1998. The end of cheap oil. *Scientific American* **278**(3), 78–83.

Central Intelligence Agency, US. 2011. *The World Factbook*. https://www.cia.gov/library/publications/the-world-factbook/fields/2002.html.

Criqui, P. 1996. Energy and climate change: socioeconomic aspects. In *ERCA – Volume 2: Physics and Chemistry of the Atmospheres of the Earth and Other Objects of the Solar System* (C. Boutron, ed.). Les Editions des Physique, Les Ulis, France, pp. 277–98.

Dresselhaus, M. S. and Thomas, I. L. 2001. Alternative energy technologies. *Nature* **414**, 332–7.

Falkowski, P., Scholes, R. J., Boyle, E. *et al*. 2000. The global carbon cycle: a test of our knowledge of Earth as a system, *Science* **290**, 291–6.

Hong, S., Candelone, J.-P., Patterson, C. C., and Boutron, C. F. 1994. Greenland ice evidence of hemispheric lead pollution two milenia ago by Greek and Roman civilizations. *Science* **265**, 1841–3.

Hong, S., Candelone, J-P., Patterson, C. C., and Boutron, C. F. 1996. History of ancient copper smelting pollution during Roman and Medieval times recorded in Greenland ice. *Science* **272**, 246–9.

Kvenvolden, K. 1993. Gas hydrates – geological perspective and global change. *Reviews of Geophysics* **31**, 173–87.

National Research Council. 2010. *Towards Sustainable Agricultural Systems in the 21st Century*. National Academy Press, Washington, DC.

National Research Council. 2011. *Renewable Fuel Standard: Potential Economic and Environmental Effects of US Biofuel Policy*. National Academy Press, Washington, DC.

Press, F. and Siever, R. 1978. *Earth*. W. H. Freeman and Company, San Francisco.

Rogner, H.-H. 1986. Long-term energy projections and novel energy systems. In *The Changing Carbon Cycle: A Global Analysis* (J. R. Trabalka and D. E. Reichle, eds). Springer-Verlag, New York, pp. 508–33.

Shotyk, W., Cherburkin, A. K., Appleby, P. G., Fankhauser, A., and Kramers, J. D. 1996. Two thousand years of atmospheric arsenic, antimony, and lead deposition recorded in an ombrottrophic peat bog profile, Jura Mouintains, Switzerland. *Earth and Planetary Science Letters* **145**, E1–E7.

Zmirov, D. 1996. Some issues on health impacts of air pollution. In *ERCA Vol. 2: Physics and Chemistry of the Atmospheres of the Earth and Other Objects of the Solar System* (Boutron, J. T. ed.). Les Editions de Physique, Les Ulis, France, pp. 265–76.

24

Coda: the once and future earth

The origin and evolution of Earth involved physical processes that operate on all matter and energy in the universe. The formation of stars is a common phenomenon in galaxies, and the formation of planetary systems is a common result of star formation. Planets are extremely common throughout the universe, and the technology to detect and characterize them continues to improve. Over 700 planets had been discovered as of the end of 2011, with even more candidates awaiting confirmation. Both the sizes and masses are becoming known for an increasing number of planets, allowing a preliminary division into rocky, icy, and gaseous classes.

In our solar system, three rocky planets had the potential early on for supporting life. Venus, Earth, and Mars were all endowed with carbon dioxide atmospheres, and at least Earth and Mars received large influxes of organic materials and water. The presence of a watery ocean was a key early step toward regulating and retaining the atmosphere. The absence or early demise of an ocean on Venus is causal to its present state: with no sink for carbon dioxide in the form of carbonates, all of the carbon dioxide remained as a massive atmosphere supporting a super-greenhouse warming: perpetually too hot to ever permit liquid water to exist.

The evidence is compelling that ancient Mars had a milder, wetter climate than today's. Mars' small size meant that gases could easily be lost to space during impacts, and plate tectonics could not be sustained. The absence of plate tectonics meant no recycling of carbon dioxide, and hence carbonate formation permanently locked up carbon dioxide in the crust. The loss of carbon dioxide progressively cooled the surface and atmosphere until liquid water froze completely. Additional loss of atmosphere by impacts completed the thinning of the Martian environment. Mars went cold more because of its size than its distance from the Sun. Venus went awry through overheating: its closer proximity to the Sun relative to Earth's distance made runaway loss of an early ocean seemingly unavoidable. Venus was left with a massive carbon dioxide atmosphere, which through the greenhouse effect cooks the surface to above the melting point of lead. However, the gradual brightening of the Sun puts the runaway point ever closer to the orbit of our own planet, and there may come a day when our planet will suffer the fate of Venus.

On Earth, the abundance of water and initiation of plate tectonics set up a relatively stable environment in which the formation of life took place. Early life had little effect on the environment, but once sufficient biomass accumulated and photosynthesis became an important biological process, the build up of oxygen profoundly changed the evolution of the surface–atmosphere system of Earth. The advent halfway through Earth's history of oxygen as the major active gas in the atmosphere enabled aerobic respiration to take place, a powerful energy source that drove the flowering of species toward increasing diversity and biomass. This flowering was delayed and then modulated over the past two billion years by climate changes induced by – among other possible causes – large impacts, orbital variations, plate tectonic movement of continents, and variations in the pace of geologic activity.

The emergence of humankind was favored by an unusual climatic epoch in which ice ages and warmer periods oscillated on 10,000- to 100,000-year cycles. Human beings, in the past 10,000 years of stable interglacial climate, have developed technologies that have enabled an explosive population growth, which now strains resources, is changing the chemistry of the atmosphere, and possibly altering the climate. In the end, these effects will be small compared to past changes over geologic timescales, but they are so sudden that they pose severe strains on the delicate balance between our civilization and the natural world upon which we depend.

The history of Earth is unique to our solar system, but it involves physical and chemical processes that are seen to operate on the other planets. Given similar initial conditions, a subset of planets orbiting other stars should have the potential for stable Earth-like conditions. The greatest uncertainty lies in the origin of life: is it the result of a set of common physical and chemical processes that could work on any number of suitable planetary surfaces? There is weak evidence that such is the case, but it is weak indeed: we know far less about how life formed than how Earth has evolved over time.

If indeed life arose in planetary systems elsewhere in the universe, intelligence may have arisen elsewhere as well. Interstellar travel is daunting from the standpoints of time and energy, but communication with other technical civilizations is possible. How many exist now depends on the lifetime of such

civilizations: we have had radio telescopes capable of interstellar communication for 50 years out of the 4.5 billion years Earth has existed. If our civilization lasts no more than 100 years, and is typical, the chance of different planetary civilizations overlapping in time is vanishingly small.

In any event, intelligence is uniquely defined by how we think and perceive the world around us. Humankind may become extinct in the next million years, because the average lifetime for vertebrate species is no more than this. Other "intelligent" forms of life may subsequently arise on this Earth or in other planetary systems, but they will not be human beings. Intelligence as defined by humans might be a rather distinct specialization; equally complex mental capability exhibited by other species could be quite different in operation and outlook.

The uniqueness of human beings at this time and on this planet is speculation only, but it should lead to a profound loneliness and an urgency to clean up our act and survive. Should we do so, and should we then turn our attention to the immense cosmos from whence we came, what extraordinary experiences lie in store for our species beyond planet Earth?

INDEX

absolute chronologies 47
  solar system events 66–68
absolute dating techniques
  lack of samples for 61
  radioisotopic dating of rocks 79
absolute zero 29–30
absorption spectra 30–31
acceleration 25
accretion stage, heat produced 120
Ackerman, T. P. 166
actinide elements 19
adenine 133–134
adenosine triphosphate (ATP) 136, 157
aerosols in the atmosphere 281
Africa
  first migration by genus *Homo* 247–248
  fossil record of human origins 246–247, 249
  second migration by genus *Homo* 248–249
African Humid Period 267
age dating
  carbon-14 (^{14}C) dating 48–49, 49–50
  cross-checks and error analysis 52
  fission track dating 52
  half-life concept 47–49
  overview 47
  parent–daughter isotopic systems 48–49, 50–52
  types of chronologies 47, 47
  use of radioactive decay 47, 47–49
  use of the cratering record 68
  see also absolute chronologies; absolute dating; relative chronologies; relative dating
age of the Earth 47
  age dating overview 47
  attempts to determine 73–74
  catastrophism versus uniformitarianism 73
  cyclic nature of geologic processes 73–74
  increasing estimates of 73–74
  sedimentary (stratigraphic) record 74
  sedimentary rock formation 73
  understanding the rock record 73–74
ages in the geologic timescale 80, 79
agriculture
  history of 288–289
  link with geological processes 76
  pressures to produce more 288–289
air pollution health risks 294–295
albedo (reflectivity) 234
alchemists 18
alkali metals 19
alkenones 57
Alpha Centauri 14
alpha decay (α decay) 29
alpha particles (α particles) 18
alternative energy sources 292
Altman, S. 154–155
aluminum
  abundance in terrestrial rocks 190
  abundance in the solid planets 114–115
  production in stars 39
amberat evidence of climate change 20
amino acids 133
  chirality (handedness) of molecules 151–152
  codons 134
  in meteorites 151, 152
  synthesis in the laboratory 151, 152
ammonia, contribution to greenhouse effect 167
anaerobic metabolism 140
Anasazi civilization 267
ancient Egyptians 3
ancient Greeks 3–4, 14–15
andesites 91
  chemical relationships 192
  formation of 192–194
andesitic volcanism, locations of 193–194
angular momentum, conservation of 102
anions, arrangement in minerals 190–191
annular eclipses 11
Antarctica
  ice core records 259–261
  life in ice-covered lakes 183
  ozone hole over 284
anthropocene 283
apatite 52
aphelion 14
apoapse 14
Apollo missions 26, 26, 52, 117, 117, 123
arc volcanism 193–194, 194
archaea 145, 145
Archean eon 80, 79
  atmospheric carbon dioxide levels 166–167, 170
  atmospheric oxygen levels 166–167, 205–207
  carbon-silicate weathering cycle 169–170
  characteristics of rocks from 195–196
  chert data 166
  climate 161
  conditions during 131
  effects of atmospheric carbon dioxide 164–166, 166–167
  evidence for liquid water 161
  formation of continents 189
  formation of protocontinents 195–196

Archean eon (*cont.*)
greenhouse effect 162
implications of the faint young Sun 164–166
origins of prokaryotic life 158–159
situation at the end of 173
temperature on Earth 164–166
transition from the Hadean era 127–128
transition to the Proterozoic 189, 196–197
Ardipithecus 246
Aristarchus of Samos 3, 14, 14–15
Aristotle 3
arsenic, as substitute for phosphorus in biomolecules 139
artificial life, as silicon-based life 139
asteroid belt 4, 6
asteroids 6
moons of 5
asthenosphere 87–89
astrology 3
astrometry 107
astronomical units (AU) 4, 14–15
astronomy, early views of the cosmos 3–4
asymptotic giant branch (AGB) stars 40
Atlantic Ocean, opening of 233
atmosphere (Earth)
and the greenhouse effect 162–164
cloud formation 163–164
convection currents 163
declining carbon dioxide abundance 170
effects of chlorofluorocarbons (CFCs) 284
history of 170
mechanism of carbon dioxide removal 167–168
origin of 125–126
ozone layer depletion 284
photochemistry 203, 203
reservoir of carbon 168
weather generation 163–164
atmosphere–ocean global circulation models (AOGCMs) 276
atmospheres on planets and moons 5
atomic nucleus
discovery of 18–20
quarks 27
understanding of workings 35
atomic number 18
periodic table 18–20
atomic radius 190
atomic weight 19, 20–21
atoms 18
discovery of nature of 18–20
Australopithecines 246–247, 247
Australopithecus afarensis 246, 247
autocatalysis 152–154
automotive emissions, health risks 295

Bacon, Sir Francis 83
bacteria 145, 211
and the origin of eukaryotes 211–212
appearance on Earth 131
cyclic photosynthesis 136
fermentation 136
banded iron formations 205–206, 209–210, 210
Barron, E. 237
basalt formation, partial melting of the mantle 191–192
basaltic crust formation 192
basalts
chemical relationships 192
formation of granites from 192–194
P-wave velocity 192
typical elemental abundances 189–190
bases in nucleic acids 133–134, 134
batholiths 195
Becquerel, Henri 74
Benner, Steven 140
Bermuda, snail fossil records 218–220
beryllium
p–p chain fusion process 36
production in stars 39, 41
beta decay (β decay) 29, 40, 40
Big Bang 16, 16, 16, 17, 99
production of helium 38
production of hydrogen 38
production of lithium 38
bilayer membranes, formation of 152–153
biofuels 292
biological effects of the K/T impact 225
biomass carbon reservoir 168
biosphere on Earth, finite life of 185–186
black holes 35, 40
black bodies 30, 31
boron
p–p chain fusion process 36
production 41
Brahe, Tycho 13
British Petroleum *Deepwater Horizon* disaster (2010) 291
Broecker, W. S. 267–268
Burgess Shale 220, 221

Cairns-Smith, A. 138
calcite structure 190
calcium carbonate, creation by shell-forming organisms 167–168
Callisto (moon of Jupiter) 5, 141, 141
Cambrian period
Burgess Shale 220, 221
diversification of life 223
Cambrian revolution 227 *see also* Ediacaran–Cambrian revolution
Cameron, A. G. W. 102
Campbell, I. E. 199–200
Cann, R. 249
carbohydrates 133
carbon
abundance in the cosmos 138
bonding properties 138–139
fusion in stars 39, 39
inorganic reservoir 168
isotopes in seafloor sediments 55
requirement of life 138–139
reservoirs on the Earth 168
stable isotopes 55
carbon-14 (^{14}C) dating 48–49, 49–50
carbon cycle
burial of carbon from organisms 204
carbon reservoirs on the Earth 168
effects of fossil fuel burning 291–292
influence of life 169
recycling of buried sediments 204
role of plate tectonics 167–168
carbon dioxide
abundance in the Archean eon 164–166, 166–167
as a greenhouse gas 164

declining atmospheric abundance 170
effect on global temperatures 271–273
evidence from paleosols 166–167
greenhouse effect in the Archean 164–166, 166–167
in the atmosphere of Venus 170
mechanism of removal from the atmosphere 167–168
projections for increase 272
uptake by living organisms 167–168
carbon dioxide cycling, carbon-silicate weathering cycle 167–168
carbon dioxide greenhouse, limits on Mars 181–182
carbon–nitrogen–oxygen (CNO) fusion cycle 37, 39
carbon-silicate weathering cycle 167–168
during the Archean eon 169–170
history of Earth's atmosphere 170
negative feedbacks 168–169
carbonaceous chondrites 51–52, 105–106
contribution to Earth's water 125
elemental abundances 189–190
see also chondritic meteorites
Cassini–Huygens mission 115, 142, 147
catalysis 152
RNA as biological catalyst 154–155
catastrophic models of continental movement 83
catastrophism versus uniformitarianism 73
cations, arrangement in minerals 190–191
Cech, T. 154–155
cells 135–136
early eukaryotes 211
essential requirements 156–158
nucleus 135–136
organelles 135–136
structure of eukaryotic cells 135–136
structure of prokaryotic cells 135
Celsius scale 29–30
Cenozoic era 80, 79
Cepheid-variable stars 15
chalcogens 20
chalcophile ("ore-loving") elements 121
chaos, definition 241
Charon (moon of Pluto) 106, 125
cheetah, genetic similarity 218
chemical bonding
and electron number 20
properties of elements 18–20
chemisynthesis 136–138
Chernobyl nuclear accident (1986) 292
cherts
as climate indicators 59
data from the Archean eon 166
Chicxulub crater, proposed K/T impact site 225–227
chirality (handedness)
and function of biological molecules 151–152
and the origins of life 155–156
chlorine, effects on the ozone layer 284
chlorofluorocarbons (CFCs)
effects on the ozone layer 284
greenhouse gases 164
in the atmosphere 271
chlorophyll 136
chloroplasts 135–136, 211
origin of 211–212
chondritic meteorites 122
constituents of 114, 114
elemental abundances 189–190
see also carbonaceous chondrites
chromosomes, in early eukaryotes 211
Chyba, Chris 167
citric acid cycle 136
civilizations, decline related to climate 267
cladograms 220–221
clathrate hydrates in seafloor sediments
climate
and Earth's movement 11
effects of continental movements 233
in the Cretaceous 235–237
in the Tertiary 237–239
influence of Earth's tilt and orbit 239–241
influences on 11
link with plate tectonics 94–95
oceanic–atmospheric connection 267–268
oscillations in the Pleistocene 239–241
role of the oceans 281–283
versus weather 280–281
climate change
African Humid Period 267
amberat evidence 20
and decline of civilizations 267
end of the last ice age 268
ice core records 259–261
influence of solar activity 267
Little Ice Age (sixteenth to nineteenth century Europe) 267
Medieval Warm Period 267
packrat midden evidence 262–264
plant pollen evidence 261–262
present climate 268
present global warming in perspective 259
tree ring evidence 264–266
variability in the Late Holocene 266–267
Younger Dryas 267–268
see also global warming; human-induced global warming
climate indicators
alkenones 57
carbon isotopes 55
cherts 59
hydrogen isotopes 56–57
nitrogen isotopic ratios 57
oxygen isotopes 55–56
sulfur isotopic ratios 57
use of stable isotopes 55, 55–57
climate models 273–276
atmosphere–ocean global circulation models (AOGCMs) 276
basic physics of the greenhouse effect 273–274
complicating factors 274–275
Cretaceous climate 237
general circulation models (GCMs) 275–276
predicted effects of global warming 276–280
role of the oceans 281
shutdown of the ocean circulation 283
climate system
negative feedbacks in 235
positive feedbacks in 234
clones 220
clouds
condensation and evaporation 274
formation of 163–164
CNO (carbon–nitrogen–oxygen) fusion cycle 37, 39
coal as a fuel 289–291

coccoliths
uptake of atmospheric carbon 55
uptake of oxygen isotopes 56
codons 134
comets 6, 106
impacts 70
Jupiter family short-period comets 106
materials brought to Earth by 125–126
Shoemaker–Levy 9 225
computer modeling *see* climate models
condensation and latent heat 163–164
conservation of energy 29–30
continental crust 90
diagnosing history and origin 190
presence of rare earth elements 190
upper and lower sections 196, 197
continental drift theory (Wegener) 83–84
continental movements
and ice ages 234
effects of 233
effects on climate 233
effects on ocean currents 233
mountain building 231, 233
supercontinent cycle 231, 231–233
volcanism 233
continental rocks 74
features of Archean rocks 195–196
first stable continental rocks 127–128
model of granite formation 194–195
continents
and sea-level changes 200
area in the Proterozoic 196–197
changing geochemistry in the Proterozoic 196–197
formation in the Archean 189, 195–196
influence on tidal effects 200–201
continuum spectra 30–31
convection currents in the atmosphere 163, 274
Copernican model of Earth motion 11
Copernican Revolution 4
Copernicus, Nicolaus 4
coral reefs 167
Corot mission 107–108, 108
Cosmic Background Explorer satellite 16
cosmic microwave background radiation 16, 17
cosmic rays 41, 49
cosmological constant (Einstein) 17
cosmos, history of 99–100
covalent bonding 20
cratering
absolute chronology of solar system events 66–68
causes of 61
impactors through time 70
on planetary bodies with atmospheres 68–69
relative age dating 61, 63–66
use to date planetary surfaces 68
see also impacts
cratering process 61–62
impact speeds 61
multiring basins 62
shock waves 61–62
craters
Meteor Crater, Arizona 6
on Mars 178
on Venus 176
Cretaceous climate 235–237
computer modeling 237
constraints on climate 235–236
evidence for climate pattern 235–236
galactic effects 237
greenhouse heating 236–237
ocean current effects 236–237
orbital variation effects 237
plate tectonic effects 236
solar output effects 237
water vapor and cloud cover 236–237
Cretaceous fossils 235–236
Cretaceous–Tertiary extinction 223–227
biological effects of the impact 225
evidence for an impact 224
interpretation as an impact event 224–225
iridium in boundary sediments 224, 225
link to Chicxulub crater 225–227
properties of boundary sediments 224
Croll, J. 240
Cro-Magnon people 252, 252, 288
Cruzan, Paul 283
Curie point 84–86, 87
Curie, Marie 74
cyanobacteria in the early oceans 206
cyclic photosynthesis 136
cytoplasm 135
cytosine 133–134

Dalton, John 18
dark energy 16–17, 17
dark matter 17
Darwin, Charles, *The Descent of Man* (1871) 246
Darwinian evolution 217, 217–218
and definition of life 131–133
dating *see* age dating
Davies, Paul 146
day length on Earth, alteration over time 200–201
de Duve, C. 157
decay, uptake of oxygen 204
Deccan Traps lavas, India 233
Deep Impact probe (USA) 106
deforestation 292
Democritus 18
density of the planets 113, 113–114
Denton, G. H. 267–268
Des Marais, David 210
deterministic chaos 241
deuterium 20–21
deuterium fractionation, climate indicator 56–57
deuterium-to-hydrogen ratio, atmosphere of Venus 174, 175
dew point 274
diatomic compounds 19
dikes (igneous rock intrusions) 77
dinosaurs (Archosauria)
Cretaceous–Tertiary extinction 223–227
fossil record 80
diorites
chemical relationships 192
formation of 193–194
distances
beyond the galactic neighborhood 15
Earth–Moon 14
Earth–Sun 13–14, 14

scientific notation 9
to nearby galaxies 15
to nearby stars and planets 14–15
to the farthest edge of the universe 15–17
to the planets 13–14
use of Cepheid-variable stars 15
DNA (deoxyribonucleic acid) 145
evolution after RNA 154
evolution of 134, 154, 157–158
in early eukaryotes 211
in eukaryotic cells 135–136
in prokaryotes 135
mitochondrial 135, 211–212
mitochondrial DNA analysis 249
nuclear 135
replication and mutations 134–135
role in protein synthesis 134
structure and replication 133–134
Doppler shift 15–16
star spectra 107
dry convection in the atmosphere 274
dwarf planets 4, 5 *see also* Eris; Pluto

Earth
cooling trend 170
Copernican model of motion 11
day length alteration over time 200–201
decreasing atmospheric carbon dioxide 170
determination of size of 3
distance from the Moon 14
distance from the Sun 13–14, 14
exchange of material with Mars 183–184
finite life of the biosphere 185–186
Goldilocks view 170
gravitational interactions with the Moon 11, 200–201
impacts from space 6
importance of the carbon cycle 170
influences on climate 11
motions in the cosmos 9–13
slowing of rotation over time 200–201
spherical nature of 3
terrestrial planet 4
uniqueness in the solar system 99
see also terrestrial planets
Earth age *see* age of the Earth
Earth axial tilt 11
influence on climate 239–241
precessional cycle 239–241
stabilizing effect of the Moon 241–242
Earth-centered cosmos concept 3–4
Earth formation
accretion process 74
basaltic crust formation 192
conditions during the Archean eon 131, 164–166, 166–167
conditions in the faint-young-sun era (Archean) 164–166, 166–167
distribution of elements 121–122
earliest evidence of life 131
early differentiation after accretion 121–122
effects of gravitational contraction 74
first stable continental rocks 127–128
formation of the Moon 123–125
from Hadean era to Archean eon 127–128
generation of the magnetic field 123
Hadean era 113
heat produced during accretion 120
historical influence of liquid water 170
information from meteorites 125
iron core formation 123
Late Heavy Bombardment 126–127
magma ocean stage 121–122
materials from impacts 125–126
origin of the atmosphere 125–126
origin of the ocean 125–126
origin of the organic reservoir 125–126
past temperature determination 55, 55–57
perspective on early history 227
perspective on history and future 299–300
radioactive heating effect 122
situation at the end of the Archean 173
source of Earth's water 125–126
timescale for early events 125
Earth orbital period 14
influence on climate 239–241
Earth structure
constituents of 114–115
constituents of the core 118–120
geologic differences to Venus 176–178
geologic history 81
internal structure 117–120
magnetic field reversals 84–87
mantle heat flow 122
radioactive element abundances 122
stratified structure 114–115
structure of the core 117–118
earthquakes
and the structure of the Earth 117–118
association with ocean ridges and trenches 84
association with subduction zones 87
P-waves 117–118
S-waves 117–118
eccentric planetary orbits 6
eclipses
annular eclipse 11
prediction of 11, 11–12
regression of the nodes 11, 11–12
see also lunar eclipse; solar eclipse
ecliptic plane 11, 11, 14
economically important minerals
Ediacaran–Cambrian revolution 220–223
absence of predators 222
as geological artifact 222–223
beginnings in the Ediacaran 221–222
carbon burial 222
causes 222–223
establishment of basic body plans 221
genetic complexity 222
near-global glaciations 222
oxygen levels 222
phylogeny 220–221
sulfide ocean 222
taxonomy 220
why it has not been repeated 223
Eemian interglacial 259–261
Einstein, Albert
conversion of mass to energy 29, 35
cosmological constant concept 17
general relativity theory 26–27
El Niño phenomenon 282–283

Eldredge, N. 218–220
electric fields 27
electromagnetic force 27
electromagnetic spectrum 30–31
electromagnetism, photons 25
electrons 18–20
  behavior of 21
  chemical bonding 20
  energy levels 20–21
  mass 18
  quantum mechanics 20–21
  wave patterns (wavefunction) 21
element production
  and life 41
  in the Big Bang 38
  l process 41
  nonstellar 41
element production in stars 25, 35–38, 38–39
  neutron removal 40
  p process (proton capture) 39, 40
  r process (rapid neutron capture) 40
  s process (neutron capture) 39–40
  supernovas
elements 17
  abundances in terrestrial rocks 189–190
  abundances in the Sun 31–33
  chalcophiles ("ore-loving") 121
  chemical bonding properties 18–20
  components of 18
  creation of artificial elements 28
  discovery of 18
  distribution in the Earth 121–122
  formation of 17
  ionic radius 190–191
  lithophiles ("rock-loving") 121
  origin of 35
  periodic table 18–20
  properties of 18–20
  siderophiles ("iron-loving") 121
elliptical orbits 13–14, 14
emission spectra 30–31
empirical models 14
enantiomers of amino acids 151–152
Enceladus (moon of Saturn) 147
energetic processes of life 136, 136–138
energy
  and mass 35
  and matter 29
  and work 29
  conservation of 29–30, 29
energy resources 289–292
  alternative sources 292
  biofuels 292
  challenges of fossil fuels 291–292
  energy use in the future 292
  fossil fuels 289–291
  fusion reactors 292
  geothermal energy 292
  hydroelectric power 292
  nuclear fission 292
  solar energy 292
  wood as fuel 292
energy-storing phosphate bonds 140
energy transfer *see* thermodynamics
ENSO (El Niño/Southern Oscillation) 282–283
enstatite ($MgSiO_3$) 114
entropy 149–151
enzymes 133, 152, 157–158
Eocene epoch 237
  age of mammals 238–239
  extinction events 237
eons in the geologic timescale 80, 79
epochs in the geologic timescale 80, 79
equilibrium in a system 149–151
eras in the geologic timescale 80, 79
Eratosthenes 3
Eris 5, 6
eubacteria 145
eukarya 145–146
eukaryotic cells 135–136
eukaryotic life 145–146
  and rise in oxygen levels 211–212
  appearance in the mid-Proterozoic 211
  appearance of complex multicellular organisms 212
  conditions for development of 211–212
  origin of 211–212
Europa (moon of Jupiter) 99, 140–142
eutectic solutions 119
evolution of complex life, possible mechanisms 215
evolution of species
  and genetic mutation 217
  and natural selection 217
  classical Darwinian model 217, 217–218
  controversy over 217
  definitions 217
  evidence for 217
  molecular clocks 135
  mutation and genetic variation 134–135
  punctuated equilibrium model 218–220
  religious views on 246
  role of the genetic code 218–220
  subspecies evolution 218–220
  "survival of the fittest" 217, 217–218
  symbiosis mechanism 220
  trigger genes 217, 219
extinctions 79
  effects of the Pleistocene ice ages 242
  Eocene 237
  Pliocene 237
  see also mass extinctions
extrusive igneous rocks 192

Fahrenheit scale 29–30
faint young Sun (faint early Sun)
  alternative theory 170
  conditions on Earth 164–166
  problem of 59, 161
  reasons for 161–162
Feinberg, J. 139, 140
felsic rocks 127–128
  chemical composition 192
  formation of 192–194
Fenchel, T. 212
fermentation 136
Finlay, B. J. 212
fission *see* nuclear fission
fission track dating 52
flagellum 135

flake tectonics (delamination) 195, 199
food production
environmental costs 288–289
pressures on 288–289
force, definition 25
forces of nature 25–29
Forget, Francois 182
formations of rocks 76–77
Formisano, V. 144
forsterite (Mg_2SiO_4) 114
Forterre, P. 145
fossil fuels 289–291
carbon reservoir 168
challenges of 291–292
combustion 204–205
effects on the carbon cycle 291–292
recoverable reserves 291–292
fossil record 77–79
Cretaceous period 235–236
early aerobic organisms 207
evidence for continental movement 83
evidence of human origins 245–246, 246–247, 249
first eukaryotic life 211
petrification process 77, 77–79
punctuated equilibrium evolutionary model 218–220
use in dating sedimentary layers 79
fracking, natural gas extraction process 291
Fraunhofer, Josef 31–33
free energy 136
free radicals 211, 295
frequency 30
Froelich, P. 237
Fukushima Daiichi nuclear accident (2011) 292
fusion reactions in stars 35–38
CNO (carbon–nitrogen–oxygen) cycle 37
p–p chain (proton–proton chain) process 36–37
fusion reactors 292

gabbro
chemical relationships 192
P-wave velocity 192
Gaia hypothesis 169
Gaia satellite 14
galaxies
distances to 15, 15
intrinsic brightness 15
red shift of spectra 15–17
Galilean satellites of Jupiter 5
Galileo Galilei 4, 26
Galileo mission 117, 141
gamma decay (γ decay) 29
Ganymede (moon of Jupiter) 5, 63–66, 141, 141, 141, 141
gas energy source *see* natural gas
general circulation models (GCMs) of climate 275–276
general relativity, theory of 26–27
genes 134
genetic code 134
role in evolution of species 218–220
genetic complexity, Ediacaran–Cambrian revolution 222
genetic mutation 134–135
and evolution 217
trigger genes 217, 219
genetic variation 134–135
genetics-first approach to life's origin 152, 154–156, 156–158
genome sequencing, Neanderthal genome 251–252
genomic analysis, evidence for human origins 249
geologic dating 47, 73
geologic succession, and geological processes 76–77
geologic time, Hutton's view 73–74
geologic timescale 79–80
as a map of Earth's geologic history 81
Grand Canyon rock layers 80
geological processes
and geological succession 76–77
continental rocks 74
cyclical nature 74–76
erosion by water 74–76
erosion of sedimentary rocks 76
igneous rocks 74, 76
metamorphic rocks 76
oceanic rocks 74
weathering processes 74–76
geological unconformities 74, 76, 76, 77, 80
geothermal energy 292
giant planets 4–5
composition 115–117
density 115–117
possible sites for life 140
Giardia intestinalis (protozoan) 211
glaciations
and the Ediacaran–Cambrian revolution 222
influences on 234, 235, 235
see also ice ages
glaciers
erosion caused by 74
evidence of glacial activity 233–234
global temperatures
and CO_2 abundance 271–273
records of 272–273
global warming
energy resources 289–292
present warming in perspective 259
urban heat island effect 273
see also human-induced global warming
global warming predictions 276–280
biosphere–climate feedbacks 279
changes in climate variability 279
degree of uncertainty 276–278
difficulty of proof 280–281
global mean increase in precipitation 278
increased continental dryness 279
large stratospheric cooling 278
life in the next quarter century 280
more severe precipitation events 279
northern polar winter surface warming 278–279
regional vegetation changes 279
rise in global mean sea level 279
summer continental warming 279
uncertain regional-scale changes 279
gluons 27
Gondwana 91–93
Gould, Stephen Jay 218–220
Grand Canyon, Arizona, rock layers 80
granites
chemical relationships 192
formation from basalts 192–194
P-wave velocity 192
possible model of formation 194–195

granitic rocks, typical elemental abundances 189–190
granitoid rocks 195
granodiorites 194, 195
granulites 195
graphite, in K/T boundary sediments 224
gravitational compression of planets 113–114
gravitational contraction 74, 74
gravitational field mapping 117
gravitational force (gravity)
  and tides 26
  cause of 26–27
  definitions 25–27
  inverse-square property 26
graviton 27
Green Revolution 288
greenalite [$Fe_3Si_2O_5(OH)_4$] 166
greenhouse effect
  basic physics of 273–274
  carbon dioxide levels in the Archean 164–166, 166–167
  climate models 273–276
  complicating factors 274–275
  in the Cretaceous 236–237
  limits on Mars 181–182
  on the Archean Earth 162
  process 162–164
  Venus 170
greenhouse gases 164
  and human-induced global warming 271–273
  evidence from paleosols 166–167
Greenland ice core records 259–261
  correction to dating 271
guanine 133–134

habitable planets, search criteria 173, 242
habitable zone
  criteria for 184–185
  implications from Venus and Mars 184–185
Hadean Earth, Titan as analogue of 142
Hadean era 113
  transition to the Archean eon 127–128
half-life concept 47–49
Halley's comet 106
halobacteria 145
halogens 19
handedness (chirality)
  and function of biological molecules 151–152
  and the origins of life 155–156
Hansen, V. L. 198
Hawaiian Islands 89, 122, 192
Hawkins, Gerald 11
health risks from pollution 294–295
heat energy 29–30
helium
  discovery of 31–33
  fusion in stars 39
  p–p chain fusion process 36–37
  production in the Big Bang 38
Helmholtz, Herman von 74
Herodotus 73
Hohokam civilization 267
Holocene Climate Optimum 273
Holocene epoch
  climate records 259
  climate variability in the Late Holocene 266–267
  climatic influence on human development 268
  interglacial climate 238, 259–261
  Younger Dryas 267–268
homeostasis 133, 154
Hominidae 246
Homo erectus 247–248, 251
  species evolved from 253
Homo genus 245, 246 *see also* human origins
Homo heidelbergensis 251
Homo neanderthalensis (Neanderthals) 249–253
Homo sapiens
  evolution of 248–249 *see also* human origins
  taxonomy 246
hot spots, eruption of basalts 192
Hoyle, Fred 11–12
Hubbard, W. B. 116
Hubble, Edwin 16
Hubble Space Telescope 15, 17, 117
  view of comet Shoemaker–Levy 9 225
human development, influence of climate 268
human-induced global warming
  and carbon dioxide abundance 271–273
  and human population growth 287–288
  chlorofluorocarbons in the atmosphere 271
  climate models 273–276
  debate over 271
  detection of trends 280–281
  difficulty of proof 280–281
  effects of food production 288–289
  global temperature records 272–273
  greenhouse gases 271–273
  Holocene Climate Optimum 273
  long-term view 283
  methane levels 271
  nitrous oxide levels 271
  predicted effects 276–280
  projections for carbon dioxide increase 272
  research agencies 277–278
  role of the oceans 281–283
  upper atmosphere ozone depletion 284
  weather versus climate 280–281
  see also global warming
human origins 245
  African fossil record 246–247, 249
  appearance of sentient organisms 245
  Archaic human populations 253
  Australopithecines 246–247, 247
  Cro-Magnon people 252, 252
  evidence from genomic analysis 249
  evolution of *Homo sapiens* 248–249
  first migration out of Africa 247–248
  genus *Homo* 245, 246
  genus *Homo* appearance in Africa 247–248
  geographical origin 246–247
  Hominidae 246
  human interest in 253
  incomplete fossil record 245–246
  interaction with Neanderthals 252
  multiregional hypothesis 248–249
  Neanderthals 249–253
  perspective on origins and future 299–300
  Pleistocene climatic setting 245
  religious views on 246
  replacement hypothesis 248–249

second migration out of Africa 248–249
social and cultural implications of research 246
spread of modern humans 252, 253
taxonomy 246
vagaries of understanding 245–246
human population growth
and human-induced global warming 287–288
and resource depletion 287–288
challenge of overpopulation 295
dependence on technology 295
Hutton, James 73–74, 76
hydrocarbon aerosols, contribution to greenhouse effect 167
hydroelectric power 292
hydrogen
escape to space 203
p–p chain fusion process 36–37
production in the Big Bang 38
stable isotopes 56–57
hydrogen bomb 29
hydrogen bonding, between water molecules 139
hydrogen fusion in stars 38–39
hydrogen fusion in the Sun, variation over time 161–162
hydrogen isotopes 20–21
as climate indicators 56–57
hydrological cycle on Earth 174
hydrothermal systems on early Mars 183

ice age on Mars 235
ice ages on Earth 169–170, 170
and continental movements 234
effects on life on Earth 242
evidence for 233–234
negative feedbacks in the climate system 235
Pleistocene epoch 238, 239–241
positive feedbacks in the climate system 234
present day (Holocene) 238
triggers for 234, 235, 235
ice core records of climate 259–261
correction to dating 271
ice-forming elements, abundance in the solid planets 115
Iceland, composition and heat flow 195
igneous rocks 74, 76
chemical relationships 192
dating 79, 79
intrusions (dikes) 77
layering 76–77
impacts
and mass extinctions 223, 223–227
Chicxulub crater 225–227
comet Shoemaker-Levy 9 on Jupiter 225
difficulty of linking to extinctions 227
effects on Mars in the past 182
evidence in K/T boundary sediments 224
exchange between Earth and Mars 183–184
impactors through time 70
interpretation of the K/T boundary 224–225
Late Heavy Bombardment 126–127
lunar origin theories 123–125
see also cratering
inclined planetary orbits 6
India, collision with the Asian continent 237–238
Industrial Revolution 288
inflation phenomenon (following the Big Bang) 16, 17
Intergovernmental Panel on Climate Change (IPCC) 277–278
interplanetary dust particles (IDPs) 61, 107
interstellar medium 61
intrusive igneous rocks 192
inverse-square property of gravity 26
Io (moon of Jupiter) 141
ionic bonding 20, 190–191
ionic radius of elements in minerals 190–191
ions 31
iridium
abundance in meteorites 224
in K/T boundary sediments 224, 225
iron
abundance in terrestrial rocks 190
abundance in the solid planets 114, 114–115
banded iron formations 205–206, 209–210, 210
in the Earth's core 118–120
production in stars 39
iron carbonates, conditions for formation 166–167
iron core formation in the Earth 123
iron meteorites, constituents of 114
iron silicates, conditions for formation 166–167
Ishtar Terra (Venus), origin of 198–199
isobars (nucleons) 40
isotopes 20–21, 39–40
as climate indicators 55, 55–57
production 41
radioactive decay 21
iterative process, inferring constituents of planets 114

joule, definition 29
Jovian planets 4–5
Joyce, Gerald 131,
Jupiter 4–5
and the asteroid belt 6
comet Shoemaker–Levy 9 impacts 225
density and composition 115–117
Galilean satellites 5
Galileo orbiter 141
moons 5, 5
possible site for life 140
see also giant planets
Jupiter family short-period comets 106
Jurassic period 80,

K/T boundary 223–227
iridium in boundary sediments 224, 225
properties of boundary sediments 224
K/T boundary event
biological effects of the impact 225
evidence for an impact 224
interpretation as an impact event 224–225
link to Chicxulub crater 225–227
Kant, Immanuel 74
Kargel, Jeffrey 180
karst topography 74
Kasting, J. F. 162, 166, 185, 209, 210
Keller, Helen 287
Kelvin, Lord *see* Thomson, William (Lord Kelvin)
Kelvin scale
Kepler, Johannes 13–14
Kepler mission (NASA) 107–108, 108
Kepler's laws 26
kerogens 168
kinetic energy 29

Knauth, Paul 58–59
Krupp, Edward 12
Kuhn, W. R. 235
Kuiper Belt 5, 6, 70, 106

l process of element production 41
Lakshmi Planum (Venus), origin of 199, 199
lanthanide elements 19
Late Heavy Bombardment 66, 126–127
latent heat 163–164
Laurasia 91–93
lavas (volcanic igneous rocks) 76
Lavoisier, Antoine Laurent 18, 18
Lay, Thorne 119
life
  anaerobic metabolism 140
  and oxygen 140
  chemisynthesis 136–138
  criteria for the habitable zone 184–185
  definitions of 131–133
  elemental building blocks 41
  elemental requirements 138–139
  energetic processes 136, 136–138
  fermentation process 136
  metabolic mechanisms 136
  photosynthesis 136
  requirement for carbon 138–139
  requirement for water 139–140
  respiration process 136
  search criteria for habitable planets 242
  search for evidence beyond the solar system 184–185
  strategies for searching for 140
  see also life, possible sites
life on Earth
  and thermodynamics 150–151
  basic structure 133
  beginning of anthropogenic influences 253
  building blocks for life 151–152
  characteristics of 145–146
  chemical to biochemical evolution 156–158
  disequilibrium and entropy 150–151
  earliest evidence on Earth 131
  effects of Pleistocene ice ages 242
  essential requirements of cells 156–158
  establishment of basic body plans 221
  finite life of the biosphere 185–186
  Gaia hypothesis 169
  Goldilocks view 170
  handedness of biological molecules 151–152
  history in perspective 227
  influence on carbon cycling 169
  information exchange and replication 133–134
  origin theories 145–146, 149, 152, 152–154, 154–156, 156–158
  origins of metabolic cycles 158–159
  origins of prokaryotic life 158–159
  phylogenetic tree 145, 145
  possible seeding from Mars 183–184
  raw materials and synthesis 151–152
  requirements for 133
  role of nucleic acids 133–134
  shadow biosphere concept 145–146
  situation in the Archean eon 158–159
life on Mars
  possible history of 184
  potential abodes on early Mars 182–183
  potential to seed Earth 183–184
  search for evidence 178, 183–184
life, possible sites
  alien life forms on Earth 145–146
  atmospheres of the giant planets 140
  Enceladus 147
  features of life on Earth 145–146
  in the solar system 140–146
  interior of Europa 140–142
  Mars past and present 142–145
  meteorite ALH84001 from Mars 144
  shadow biosphere concept 145–146
  Titan 142
light-year 14, 14
limestone, erosion by water 74
lipids 133
  formation of bilayer membranes 152–153
lithium production 41
  in the Big Bang 38
  p–p chain fusion process 36
lithophile ("rock-loving") elements121
lithosphere 87–89
Little Ice Age (sixteenth to nineteenth century Europe) 267, 271, 280–281
Lovelock, James 169
Lowe, Donald 58–59
Lowell, Percival 174
Lucretius 18
Lucy (*Australopithecus afarensis*) 246
lunar eclipse 3, 9, 11, 14
  prediction 11, 11–12

mafic rocks 127–128, 192
Magellan spacecraft (US) 176–178
  radar images of Venus 198–199, 200
magmas 76
magnesium
  abundance in terrestrial rocks 190
  abundance in the solid planets 114, 114–115
  production in stars 39, 39
magnetic fields 27
  Earth's magnetic field 123
magnetic imprints on rocks 84–87
magnetite (Fe_3O_4) 144
magnetotactic bacteria 144
main sequence stars 37–38
mammals
  expansion of 238–239
  history of 223
Mann, M. E. 280
mantle 87–89
  elemental abundances 189–190
mantle heat flow, inhibition by continental masses 231–232
mantle plumes 89, 122
Margulis, Lynn 203, 211, 211
Mars 4
  atmosphere 170
  chaotic axial tilt 241
  climate at the time of the Archean 161
  conditions in the past 142–145
  conditions today 178
  constituents of 114–115
  cratering record 63

dust storms 178
early differentiation after accretion 121–122
effects of radioactive heating 122
exchange of material with Earth 183–184
formation of 113
geologic history 81
geology 178–179
heat produced during accretion 120
ice age 235
impact craters 178
inability to recycle carbonates 170
inhospitable conditions for life 178
lifeless appearance 99
liquid water in the past 170
meteorite ALH84001 144
methane release into the atmosphere 183
potential for terraforming 185
presence of frozen water 178
relative chronology 47
search for glacial features 234
search for life on 142–145, 178
SNC meteorites 61
tilt (obliquity) 178
Valles Marineris canyon 81
Viking missions 142, 144
volcanoes 178–179
see also terrestrial planets
Mars climate history
abodes for life on early Mars 182–183
absence of plate tectonics 178–179
conditions on Mars today 178
effects of impacts 182
evidence from robotic missions 173
evidence of liquid water in the past 179–180
geological indications of early warmer conditions 179–180
implications for life elsewhere 184–185
limits to a carbon dioxide greenhouse 181–182
Martian geology 178–179
possible history of Mars 184
problem of warming early Mars 181–184
search for evidence of early climate 183–184
search for evidence of life 183–184
Mars Exploration Rovers (*Spirit and Opportunity*) 145, 178, 180, 183
Mars Express Orbiter 144, 180, 180, 183
Mars Phoenix Lander 180
Mars Reconnaissance Orbiter 180, 180
Mars Surveyor mission 179
Marshall, H. 235
mass, conversion to energy 29, 35
mass extinction events
causes 223
Cretaceous–Tertiary extinction 223–227
difficulty linking to impacts 227
impact events 223, 223–227
Phanerozoic eon 223, 223–227
massive vector bosons 29
matter
conversion to energy 29, 35
microscopic constitution of 21
search for understanding of 18
Maunder Minimum 267
Mayan calendars and numbering system 12–13
McKay, Chris 133, 183
McKay, David 144
Medieval Warm Period 267
melt spherules, in K/T boundary sediments 224, 225
Mendeleev, Dmitri 18–20
Mercury 4
constituents of 114–115
crustal evolution 121–122
effects of radioactive heating 122
heat produced during accretion 120
Mesozoic era 80, 79
messenger RNA 134
metabolic cycles, origins of 158–159
metabolic mechanisms 136
similarity in aerobic eukaryotes 212
variety in bacteria 212
metabolic processes
anaerobic metabolism 140
and photosynthesis 136
energy supply 136
metabolism first approach to life's origin 152, 152–154, 156–158
metamorphic rocks 76, 192
dating 79, 79
layering 76–77
Meteor Crater, Arizona 6
meteorites 33, 105–106
ALH84001 from Mars 144
amino acids in 151, 152
constituents of 114, 114
contribution to Earth's water 125
elemental abundances 189–190
information about Earth's history 125
iridium abundance 224
Murchison meteorite 156
platinum-group elements 224
radioisotope age determination 51–52
SNC meteorites 61
see also carbonaceous chondrites; chondritic meteorites
methane (as fuel) 291, 291
methane (atmospheric)
and global warming 271
contribution to greenhouse effect 167
greenhouse gas 164
release into Martian atmosphere 183
methane hydrates in seafloor sediments
methanogens 145
Mexico City, air pollution problem 295
microlensing 108
microscopic constitution of matter 21
mid-Atlantic ridge 87
mid-ocean ridges 84, 89, 195
eruption of basalts 192
Milankovitch cycles 240–241, 268
Milky Way Galaxy 14, 15, 99–100
Miller, Stanley 151
Miller–Urey flask experiment 151, 152
mineral structure 190–191
arrangement of anions 190–191
arrangement of cations 190–191
ionic bonding 190–191
minerals
of economic importance
unstable in presence of oxygen 205
mining
economically important minerals
environmental impacts

Miocene epoch, age of mammals 238–239
mitochondria 135–136, 211
origin of 211–212
mitochondrial DNA 135
mitochondrial DNA analysis, evidence for human origins 249
Mogollon civilization 267
Mohorovičič (Moho) discontinuity 119
moist convection in the atmosphere 274
moist greenhouse effect
long-term fate of the Earth 185–186
runaway theory for Venus 174–176
molecular clocks and mutation rate 135
molecular clouds and star formation 100–101
molecules 18
monomers 133
Moon (of Earth) 5, 5
age of Moon rocks 52
and tidal patterns 200–201
constituents of 114–115
crustal evolution 121–122
distance from Earth 14, 200–201
gravitational interactions with Earth 200–201
heat produced during accretion 120
influence on Earth's axial tilt 11, 241–242
lasers reflected from 90
lunar eclipses 3, 9, 11, 11, 11–12
movement in the sky 9–13
orbit around the Earth 123
orbital plane 11
origin theories 83, 113, 123–125
phases of the Moon 11
relative chronology 47
moons in the solar system 5, 5
atmospheres 5
constituents of 114–115
possibility of sedimentary processes 81
see also specific moons
moraines created by glaciers 234
Morowitz, Harold 158–159
Mount Everest 231
Mount Pinatubo 167–168
Mount St. Helens 91, 167–168
mountain building, continental collisions 231, 233
multicellular organisms
appearance of complex organisms 212, 215
proliferation in the Phanerozoic 215
multiring basins 62
Mumma, M. 144
Murchison meteorite 156
mutation and genetic variation 134–135
mutation rate, and molecular clocks 135

natural gas energy 290–291
fracking extraction process 291
gas hydrates (clathrate hydrates) in seafloor sediments
methane reserves 291, 291
shale reserves 291
natural selection and evolution 217, 217–218
Neanderthals 249–253
climate setting 249–250
decline and extinction 252
evolution of 253
evolutionary origins 251
genome sequencing 251–252
interaction with modern humans 252
lifestyle and tools 252
physical features 250–252
near-Earth asteroids 6
negative feedbacks
in the carbon-silicate weathering cycle 168–169
in the climate system 235
neon, production in stars 39, 39
Neptune 4–5
density and composition 115–117
possible site for life 140
trans-Neptunian region 5
see also giant planets
neutrinos 17, 36
neutron stars 18, 38, 40
neutrons 18, 18, 20–21, 38
mass 18
removal from the nucleus 40
New Horizons mission (NASA) 106
Newton, Isaac 14, 25
Nice model of planetary configurations 126–127
nickel, abundance in the solid planets 114
nitriles 151
nitrogen
isotopic ratios 57
production in stars 39
nitrous oxide
and global warming 271
greenhouse gas 164
in the atmosphere 295
noble gases 20, 20
nonchiral molecules 152
North Atlantic, thermohaline circulation 267–268
northern polar winter surface warming 278–279
nuclear fission 29, 292
nuclear fusion 29
element production in stars 38–39
energy source for the Sun 74
nuclear fusion reactors 292
nuclear reactions 25
nuclear reactor accidents 292
nucleic acids 133, 133–134
handedness (chirality) of sugars 151–152
role in protein synthesis 134
see also DNA; RNA
nucleotides 133–134
nuclides 39–40
numerical values, scientific notation 9

oceans
origin of 125–126
reservoir of carbon 168
role in Earth's climate 281–283
ocean circulation
basic processes 281–282
buoyancy force 281–282
El Niño phenomenon 282–283
shutdown after prolonged warming 283
Southern Oscillation 282–283
wind stress 281–282
ocean currents
effects of continental movements 233
effects on Cretaceous climate 236–237
ocean floor

age of rocks 93
evidence for plate tectonics 84
reservoir of carbon 168
oceanic–atmospheric connection to climate 267–268
oceanic crust 74, 90
age of 231
formation 192
Oerlemans, J. 273
oil as an energy source 290–291
Oligocene epoch, age of mammals 238–239
olivine structure 190
Oort Cloud 6, 70, 106
ophiolite suites 93
orbital motion of planets 26
orbital period of the planets 14
organic molecules 133
organic reservoir on Earth, origin of 125–126
oxygen
abundance in terrestrial rocks 190
abundance in the cosmos 139
and life 196–197
and strategies to search for life 140
isotopes as climate indicators 55–56
isotopes in water 55–56
production in stars 39
stable isotopes 55–56
use by eukaryotes 135–136
oxygen (atmospheric)
and onset of eukaryotic life 211–212
and the Ediacaran–Cambrian revolution 222
balance between loss and gain 208
history of oxygen on Earth 209–210
increase in the Proterozoic 201
level in the Archean 166–167
model for the rise of oxygen 209
reservoirs of oxygen 208–209
reservoirs of reducing compounds 208–209
rise in the Proterozoic 207
shield against ultraviolet radiation 211
oxygen anion, arrangement in minerals 190–191
oxygen balance, with and without life 205
oxygen cycle 203–205
oxygen levels on early Earth 205
banded iron formations 205–206, 209–210, 210
fossils of aerobic organisms 207
limits on 205–207
minerals unstable in presence of oxygen 205
redbed sediments 206, 210, 210
oxygen revolution 203
oxygen sources and sinks on Earth 203–205
burial of carbon from organisms 204
decay 204
fossil fuel combustion 204–205
photochemistry and escape of hydrogen 203
photosynthesis 203, 205
recycling of buried sediments 204
respiration 204
volcanism 203
weathering of rock 203
ozone
health risk in the lower atmosphere 295
ozone layer depletion 284
production in the atmosphere 203
shield against ultraviolet radiation 211
p process (proton capture) 39, 40
P-waves 117–118, 192
Pacific ring of fire 233
packrat midden evidence of climate change 262–264
paleomagnetism 84–87
paleosols, evidence for carbon dioxide abundance 166–167
Paleozoic era 80, 79
Pangaea 91–93, 93, 93
break up of 233, 236
Pannotia 93
Panthalassa 91–93
parallax phenomenon 4
parallax shift 14, 14–15
Paranthropus 246–247, 247
parent–daughter isotopic systems 48–49, 50–52
parsec (parallax-second) 14
partial melting 191–192, 192–194
particulates in the atmosphere 281
health risks 295
Pathfinder Mars lander 178
Patterson, Claire 52
Pavlov, A. A. 209
peptide nucleic acids (PNAs) 155
peptides 135
periapse 14
peridotite 192, 192
perihelion 14
periodic table of the elements 18–20, 20
periods in the geologic timescale 80, 79
peritectic solutions 119
Permian mass extinction 223, 223
petrification of organic remains 77, 77–79
Phanerozoic eon 80, 79, 80
Cambrian revolution 220–223
Ediacaran–Cambrian revolution 220–223
mass extinction events 223, 223–227
place in Earth's history 227
plate tectonics 231, 231–233
proliferation of complex life 215
timescale 215
Phillips, R. J. 198
phosphate bonds
energy storage 139, 140
functions in metabolic processes 136
phosphate groups in nucleic acids 133–134
phosphorus, production in stars 39 *see also* phosphate bonds; phosphate groups
photochemistry in the atmosphere 203, 203
photon, definition 25
photons 27, 30–31
absorption by greenhouse gases 164
movement in the greenhouse effect 162–164
solar-optical type 164
spectra 30–31
terrestrial-infrared type 164
photosphere of the Sun 31–33
photosynthesis 133, 136
chloroplasts 135–136
earliest evidence for 131
oxygen produced 203, 205
photosynthesizing organisms
response to enhanced CO_2 levels 279
spread of 209–210, 210
phyllosilicates 180, 183, 183

phylogenetic trees 145, 145, 220–221
phylogeny 220–221
phylum (taxonomic level) 220–221
Pierrehumbert, Ray 182
Piltdown man hoax (1913) 246
Pioneer Venus entry probes 174
Planck, Max 30
Planck function 30
planet formation 102–103
  and disks around protostars 103–104
  effects of gravitational contraction 74
  end of 104–105
planetary surfaces, use of cratering to date 68
planetary systems, search for *see* search for planetary systems
planetesimals 120, 103, 113, 126
planets
  accretion 120
  atmospheres 5
  bulk compositions 113–117
  composition of giant planets 115–117
  composition of terrestrial planets 114–115
  density determination 113, 113–114
  differentiation 117
  distances to 13–14, 14–15
  giant (Jovian) planets 4–5
  inferring constituents of 114
  measuring mass and size 113
  moons 5, 5
  motions in the sky 11, 13–14, 14
  orbits 6, 14, 26
  possibility of sedimentary processes 81
  properties of the giant planets 6
  properties of the terrestrial planets 6
  radioactive heating 122
  ring systems 5
  rotational axes 6
  self-compression effect 113–114
  spins 6
  structure of the solar system 4–6
  terrestrial planets 4
plankton 169
plant pollen evidence of climate change 261–262
plasma 18, 31
plastids 135–136, 211
  origin of 211–212
plate boundaries 90
plate tectonics 83, 170
  absence on Mars 178–179
  after the Proterozoic 197–198
  and sea-level changes 200
  and water 199–200
  as unique to the Earth 94
  basic model 87–91
  beginning on Earth 196–197
  catastrophic models 83
  driving force 94–95
  early evidence for 83–84
  effects of continental collision and separation 233
  effects on Cretaceous climate 236
  evidence from paleomagnetism 84–87
  evidence from seafloor topography 84
  failure to take hold on Venus 176–178, 198–199, 200
  fossil evidence 83
  future predictions 93
  genesis after World War II 87
  geologic record on land 87
  historical development of 83–84
  in the early Earth 210
  in the Phanerozoic 231, 231–233
  link with climate 94–95
  locations of earthquakes 87
  modern plate tectonics 197–198
  past motions of the plates 94
  role in carbon cycling 167–168
  role of water 94–95, 193–194
  shift to modern mode 201
  speed of movement of plates 90
  subduction zones 87, 167–168
  supercontinents 94, 231, 231–233
  triple junctions 91
  Wegener's continental drift theory 83–84
Plato 3
Pleistocene epoch
  causes of ice ages 239–241
  climate variations 259, 259–261
  effects of oscillatory ice ages 242
  ice ages 238
  oscillatory nature of the climate 239–241
  setting for human origins 245
Pliocene epoch, extinction events 237
Pluto 4
  classification of 5
  constituents of 115
  moons 5, 106, 125
  NASA *New Horizons* mission 106
  orbit 6
plutonic igneous rocks 74, 76
pollution
  associated with industrial processes 294–295
  associated with mining
  ozone in the lower atmosphere 295
  potential health risks 294–295
polycyclic aromatic hydrocarbons (PAHs) 144
polymers 133
positive feedbacks in the climate system 234
positrons 36
potassium
  abundance in terrestrial rocks 190
  radioactive isotope (^{40}K) 122
potential energy 29
power, definition 29
p–p chain (proton–proton chain) fusion process 36–37
precipitation, effects of global warming 278, 279
predator–prey food chains 212
pressure-release partial melting of the mantle 191–192
Priscoan eon 80, 79, 113
Prochloron (bacterium)
prokaryotic cells 135
prokaryotic life 145–146
  origins of 158–159
protein synthesis, role of nucleic acids 134
proteins 133
  handedness (chirality) of amino acids 151–152
Proterozoic eon 80, 79, 203
  appearance of eukaryotic life 211
  changing geochemistry of the continents 196–197
  increasing oxygen in the atmosphere 201, 205–207, 207

shift to modern plate tectonics mode 201
transition from the Archean 189, 196–197
protium 20–21
protocontinents, formation in the Archean eon 195–196
protons 18–20, 18
protoplanetary disk evolution 102–103
dissipation in the nebula 102–103
formation of the nebula 102
residual static nebula 103
terminal accumulation of the star 103
protostars, disks around 103–104
Proxima Centauri 4
Ptolemy 3–4
punctuated equilibrium evolutionary model 218–220
pyrite 205

quantum mechanics 20–21, 21, 30
quarks 27

r process (rapid neutron capture) 40
racemic mixtures 152
radar mapping, surface of Venus 176
radial velocity of stars 107
radioactive decay 21, 29, 41
half-life concept 47–49
use in dating 47, 47–49
see also alpha decay; beta decay; gamma decay
radioactive heating
effects on planets 122
Europa 141
radioactive isotopes, parent and daughter measurement 48–49, 50–52
radioactivity 29
discovery of 74
radiocarbon dating 48–49, 49–50
radioisotopic dating of rocks 79
Ramsey, William 31–33
Randall, Lisa 16
rare earth elements 195
abundance in Archean rocks 195
in rocks 190
Raymo, M. 237
red dwarfs 37
red giant stars 39
redbed formations 206, 210, 210
reduced carbon 204
reducing compounds, reservoirs on Earth 208–209
reflection of photons 30
regression of the nodes 11, 11–12
relative dating, cratering record 61 *see also* geologic dating/geologic layering
relative chronologies 47
resources
alternative energy sources 292
and the growing human population 287–288
depletion of 287
energy resources 289–292
respiration 136, 204
rhyolite, chemical relationships 192
ribose molecules, chirality 155–156
ribosomal RNA 134
ribosomes 135
mitochondrial 211–212
Ricardo, A. 155
ring systems 5
RNA (ribonucleic acid) 133–134, 145
bases 134, 134
codons 134
evolution of 134
forms of 134
in early eukaryotes 211
in prokaryotes 135
mitochondrial 211–212
phylogenetic tree 145, 145
role in protein synthesis 134
structure 134, 134
RNA and the origin of life 152, 154–156, 156–158
abiotic formation of RNA 154
chirality of ribose molecules 155–156
problem of abiotic invention of RNA 155–156
RNA as biological catalyst 154–155
RNA evolution before DNA 154
role as replicator 154–155
rock classes, chemical relationships 192
rock composition, effects on *P*-wave velocity 192
rock formations 76–77
rock strata (stratigraphic section) 76–77
rock weathering
paleosols 166–167
silicate rocks 167–168
Rodinia 93–94
Röntgen, Wilhelm 74
Rosetta mission (ESA) 106
rubidium–strontium decay, measurement of 50–52
Ruddiman, W. 237
runaway greenhouse theory for Venus 174
Rutherford, Ernest 18

s process (neutron capture) 39–40
S-waves 117–118
Sagan, Carl 139, 140, 167
Sagan, D. 203
Salpeter, Edwin 116
San Andreas fault system, California 84, 87, 119, 90, 90
satellites, Earth-orbiting
Saturn 4–5
density and composition 115–117
moons 5, 5
possible site for life 140
see also giant planets
scanning tunneling microscopy 21
scientific notation 9
Scopes Monkey Trial (1925) 246
Scott, David R. 26, 26
sea level rise 279
seafloor rocks
age of 93
magnetic orientations 84–87
seafloor sediments
carbon isotopes 55
coccoliths 55, 56
gas hydrates (clathrate hydrates) in
oxygen isotopes 55–56, 56
seafloor topography, evidence for plate tectonics 84
sea-level changes, and plate tectonics 200
search for planetary systems 107–110
astrometry 107
criteria for habitable planets 242
direct techniques 110

search for planetary systems (*cont.*)
indirect techniques 107–110
microlensing 108
radial velocity of stars 107
use of transits 107–108
see also habitable planets; habitable zone
sedimentary processes, on planets and moons 81
sedimentary (stratigraphic) record 74
Grand Canyon, Arizona 80
sedimentary rocks 73, 76–77, 192
dating using fossils 79
formation of 74–76
geologic cycle 74–76
lithification process 73–74
loss of layers to erosion 76
seismometers 87
Seno, Nicholas 73
serpentinization 183
shadow biosphere concept 145–146
shales, natural gas reserves in 291
Shapiro, R. 139, 140
shell-forming organisms
creation of calcium carbonate 167–168
influence on carbon cycling 169
shocked quartz, in K/T boundary sediments 224, 224
Siccar Point (Scotland) geological unconformity 74
siderite ($FeCO_3$) 144, 167
siderophile ("iron-loving") elements 121
silica, content of igneous rocks 192 *see also* cherts
silicate minerals, abundance in the solid planets 114, 114–115, 115
silicate rocks, weathering process 167–168
silicon
abundance in terrestrial rocks 190
as potential basis for life 138–139
bonding properties 138–139
fusion in stars 39, 39
in artificial life 139
SNC (Shergottites–Nakhlites–Chassigny) meteorites 61
snowball Earth episodes 58, 233, 234
sodium
abundance in terrestrial rocks 190
production in stars 39
soil-forming microorganisms 169
solar activity, influence on climate change 267
solar eclipse 9, 11, 161
prediction 11, 11–12
solar energy 292
solar nebula evolution 102–103
dissipation in the nebula 102–103
formation of the nebula 102
residual static nebula 103
terminal accumulation of the star 103
solar system
age determination 51–52
early models 3–4
giant (Jovian) planets 4–5
moons 5, 5
movements of solar system objects 9–13
orbits of the planets 6
planetary rotational axes 6
planetary spins 6
possible sites for life 140–146
properties of the planets 6
structure of 4–6
terrestrial planets 4
solar system formation
end of planet formation 104–105
history of the cosmos 99–100
planet formation 103–104
primitive material present today 105–107
protoplanetary disk evolution 102–103
star formation 100–105
unique properties of the Earth 99
solar wind 33, 170, 103
solid planets, constituents of 114–115
solid state convection 120
sonar technology, seafloor mapping 84
Southern Oscillation 282–283
space, expansion of 15–17
space–time, relativity theory 26–27
species concept, definitions 217
spectra of photons 30–31
spectrometers (spectrographs) 15–16, 30
speed 25
spiral galaxies 15
stable isotopes, carbon 55
star formation 100–105
birth of a star 101–102
conservation of angular momentum 102
disks around protostars 103–104
end of planet formation 104–105
formation of planets 102–103
giant molecular clouds 100–101
start of 101
star spectra, Doppler shift 107
Stardust probe (USA) 106
stars
creation of elements 35–38, 38–39
distances to 14–15
effects of gravitational contraction 74
element production processes 39–40
factors affecting final fate 39
fusion reactions 35–38
intrinsic brightness 15
nuclear reactions 25
stellar main sequence 37–38
stellar nucleosynthesis 38
Stevenson, David 116–117
Stonehenge (Salisbury Plain, England) 11–12
strand lines 234
strata (stratigraphic section) 76–77
stratigraphic record 74
stratosphere 174–175
stratospheric cooling 278
stromatolites 131, 158, 196
strong nuclear force 27–28
subduction 89–90, 167–168
earthquakes associated with 87
possible origin of 195–196
role of water 193–194
subspecies
definition 218
evolution of 218–220
sugar groups in nucleic acids 133–134, 134
sugars 133
production during photosynthesis 136
sulfide ocean stage 222
sulfothermophiles 145

sulfur
isotopic ratios 57
production in stars 39
summer continental warming 279
Sun
absorption spectra 31–33
abundances of elements in 31–33
age determination 52
alternative to the faint early Sun theory 170
collapse to a white dwarf 39
distance from the Earth 13–14, 14
estimates of age of 74
faint early Sun (faint young Sun) 59
increase in temperature over time 161–162
luminosity in the Archean eon 164–166
movement in relation to the Earth 3–4
movement in the sky 9–13
nuclear fusion energy source 74
nuclear reactions 25
photosphere 31–33
rate of hydrogen fusion over time 161–162
solar eclipses 9, 11, 11, 11–12, 161
transit of Venus 14–15
variation in luminosity over time 161–162
variation in output 161
Sun evolution, consequences for life on Earth 185–186
sunlight, understanding of origin 35
sunspot activity and climate 267
supercontinents 94, 210, 231, 231–233
supernovas 17, 40
element formation 39
Type 1A 15
"survival of the fittest" evolutionary model 217, 217–218
system equilibrium 149–151

taxonomy 220
Taylor, S. R. 199–200
technology, human dependence on 295
temperature
definition 29
measurement 29–30
scales 29–30
terraforming of Mars 185
terrestrial planets 4, 113
Tertiary period 80, 79, 237–239
Tethys seaway 91–93
thermodynamics
and life 150–151
second law of 149–151
thermohaline circulation, North Atlantic 267–268
Thermoplasma (bacterium) 211
Thomson, William (Lord Kelvin) 35, 74
thorium, radioactive isotope (^{232}Th) 122
Three Mile Island nuclear accident (1979) 292
thymine 133–134
Tibetan Plateau 231, 237–238
tidal heating in Jupiter's moons 141
tidal patterns and day length 200–201
tidal wave action, evidence in K/T boundary sediments 224, 224
tides 26
time concept, linear and cyclical aspects 12–13
Titan (moon of Saturn) 5, 5, 81, 99, 141
as analogue of the Hadean Earth 142
atmosphere 151, 167
Cassini–Huygens mission 142
possible chemistry of life on 140
possible site for life 142
Tombaugh, Clyde 5
Toon, Brian 182
Toon, O. B. 167
transfer RNA 134
transform faults 90
transit of Venus 14–15
transits, use in search for planets 107–108
trans-Neptunian region 5
tree rings, evidence of climate change 264–266
triatomic compounds 20
trigger genes 217, 219
triple junction of tectonic plates 91
tritium 20–21
Triton (moon of Neptune) 5
tropopause 174–175, 274
troposphere 174–175
Tully–Fisher relation 15
Type 1A supernovas 15

UK37 Index 57
ultraviolet protection, ozone layer depletion 284
ultraviolet radiation, ozone shield in the stratosphere 211
unconformities *see* geological unconformities
uniformitarianism versus catastrophism 73
universe
expansion of 15–17
measuring the size of 15–17
uracil 134, 134
uraninite 205, 209
uranium
deposits in Proterozoic sediments 196
radioactive isotopes 122
Uranus 4–5
density and composition 115–117
planetary spin 6
possible site for life 140
see also giant planets
urban heat island effect 273
Urey, Harold 151
US National Research Council 277–278

valence 19
velocity 25
Vendian period *see* Ediacaran period
Venera Soviet Venus probes 176, 176
Venus 4, 12
carbon dioxide in the atmosphere 170
constituents of 114–115
deuterium-to-hydrogen ratio 174, 175
early differentiation after accretion 121–122
effects of radioactive heating 122
failure of plate tectonics 176–178, 198–199, 200
formation of 113
geologic differences to Earth 176–178
greenhouse effect 163, 170
heat produced during accretion 120
impact craters 176
inability to produce carbonates 170
lack of a moon 125
lack of horizontal crustal movement 198–199
origin of Ishtar Terra 198–199

Venus (*cont.*)
origin of Lakshmi Planum 199, 199
planetary spin 6
radar mapping of the surface 176
transit of the Sun 14–15
see also terrestrial planets
Venus' climate history
evidence from robotic missions 173
implications for life elsewhere 184–185
loss of surface water 173–176
moist greenhouse runaway theory 174–176
runaway greenhouse theory 174
surface of Venus 176–178
thick carbon dioxide atmosphere 173–176
vesicle approach to life's origin 152, 152–154, 156–158
Viking missions 117, 142, 144, 144, 178, 178
virons 135, 145
viruses 135, 145, 211
visible light spectrum 27, 30–31
volcanic igneous rocks 74, 76
volcanism
and continental collision and separation 233
and extinction events 225
association with ocean ridges and trenches 84
deep-sea volcanic vents 136–138
extrusive igneous rock types 192
gases produced 203
Hawaiian island chain 122
Mount St. Helens 91
on Mars 178–179
release of carbon dioxide 167–168
Voyager missions 117, 117, 141, 141

Wahlen, M. 271
Walker, J. C. G. 168–169, 235
water
abundance in the solar system 139
and plate tectonics 199–200
erosion of rocks 74–76
existence as liquid 139–140
hydrogen bonding between molecules 139
influence on Earth's atmosphere 170
liquid interior of Europa 140–142
oxygen isotopes in 55–56
potential alternatives in biological systems 140
presence on Mars 178
properties of 139–140
requirement of life 139–140
role in partial melting 192–194
role in plate tectonics 193–194
role in subduction 193–194
source of Earth's water 125–126
water clouds on the giant planets 140
water cycling, role of plate tectonics 193–194
water ice, on moons and Pluto 115
water vapor, as a greenhouse gas 164
watt, definition 29
wavelength 30
weak nuclear force 29
weather
generation of 163–164
versus climate 280–281
weathering processes 74–76, 203
Wegener, Alfred 83–84
white dwarf stars 39
Whitmire, D. 170
Wilson, A. 249
Wilson, J. Tuzo 232
Wilson Microwave Anisotropy Probe 16
Wolf, E. T. 167
wood as fuel 292
work, and energy 29

Younger Dryas 267–268

zircon 52
zodiacal light 6

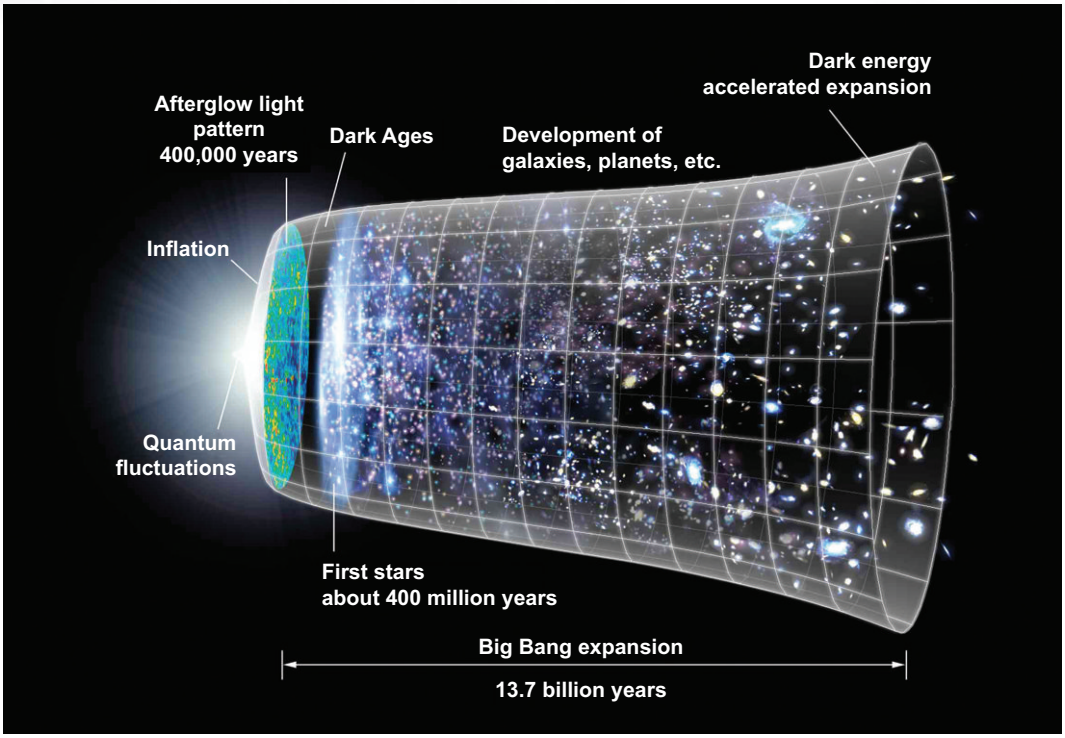


Figure 2.5 Schematic history of the entire cosmos, in which time flows from left to right. Immediately after the Big Bang all of the cosmos is dominated by fluctuations on a quantum scale, and coherent macroscopic reality as we know it does not exist. Then the scale of the universe greatly expands, in a phenomenon known as inflation, leading to the afterglow light pattern of fluctuations in the "cosmic microwave background" radiation that we see today. As the first stars form, some 400 million years after the Big Bang, formation of elements (Chapter 4) begins. Expansion of the cosmos appears to be under acceleration today, associated with a repulsive "dark energy" whose nature is not understood. Figure courtesy NASA WMAP Science Team.

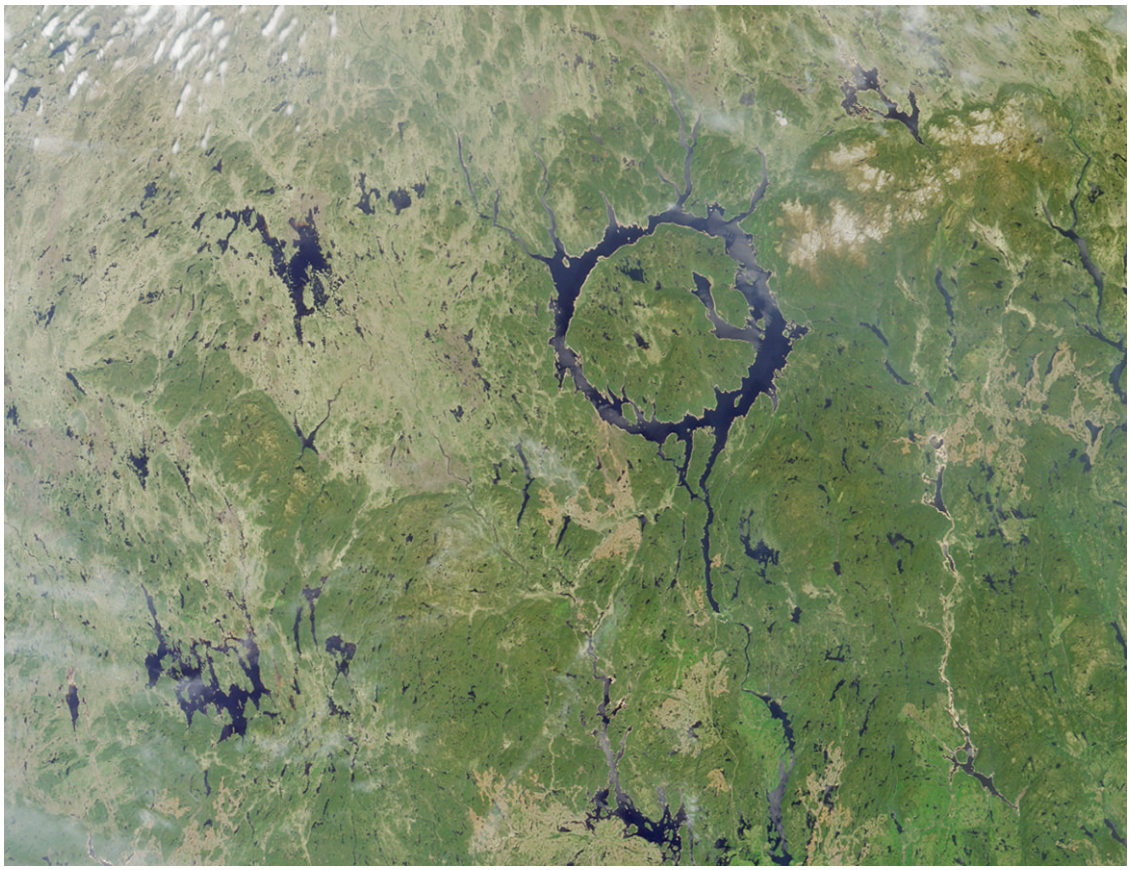


Figure 7.3 Varieties of impact craters: (a) large multiring basin, Mare Oriental, on the Moon; (b) classic large crater, Copernicus, with central peak, on the Moon; (c) smaller lunar craters without peaks; (d) pedestal crater on Mars, formed by melting of ground ice during impact. (e) relaxed craters, or *palimpsests* on Ganymede (*Voyager* image); (f) eroded crater on Earth, now comprising Lake Manicouagan, Quebec, Canada. Photos (a) through (f) are courtesy of NASA.

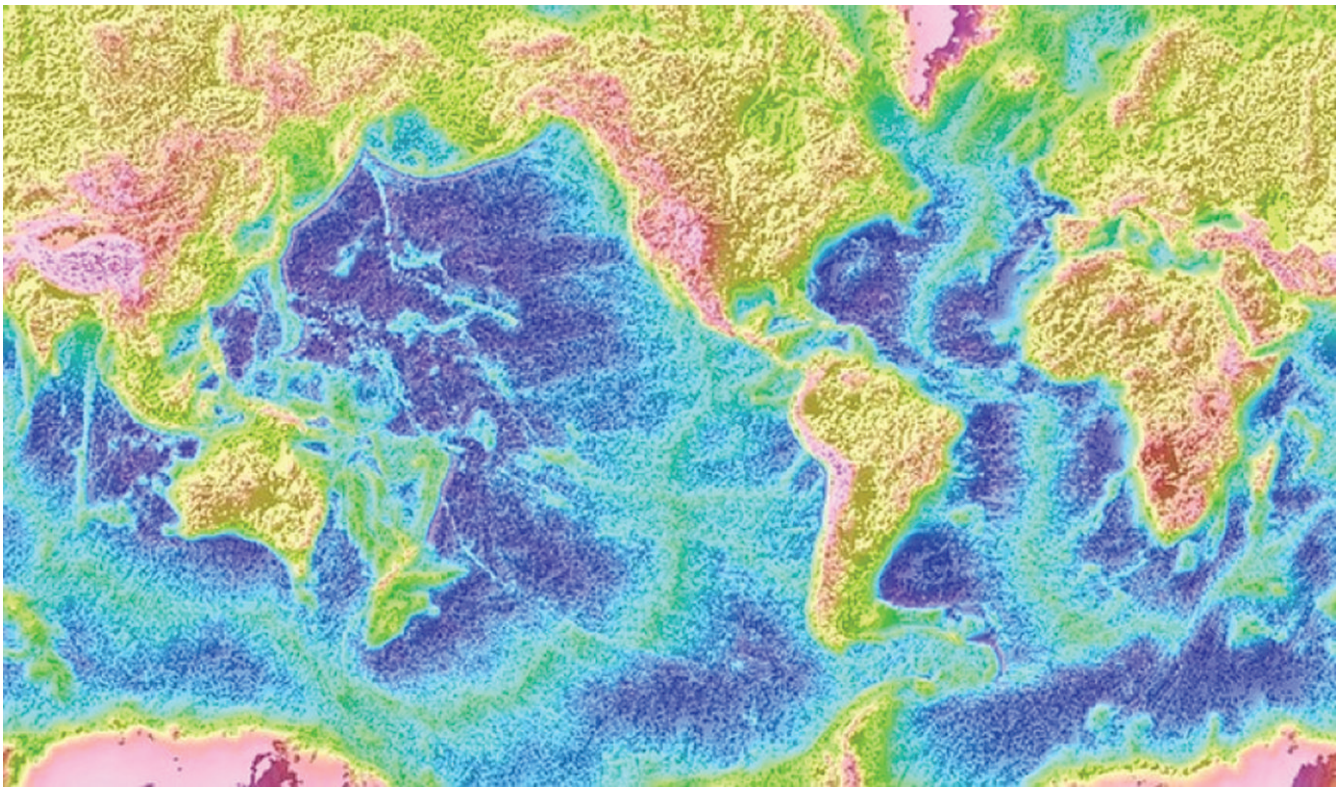


Figure 9.2 Map of the Earth's topography. Beneath the oceans, the seafloor topography is dominated by vast mid-ocean ridges, transform faults, and subduction zones, as described in the text. Red is highest, and blue is lowest, elevation.

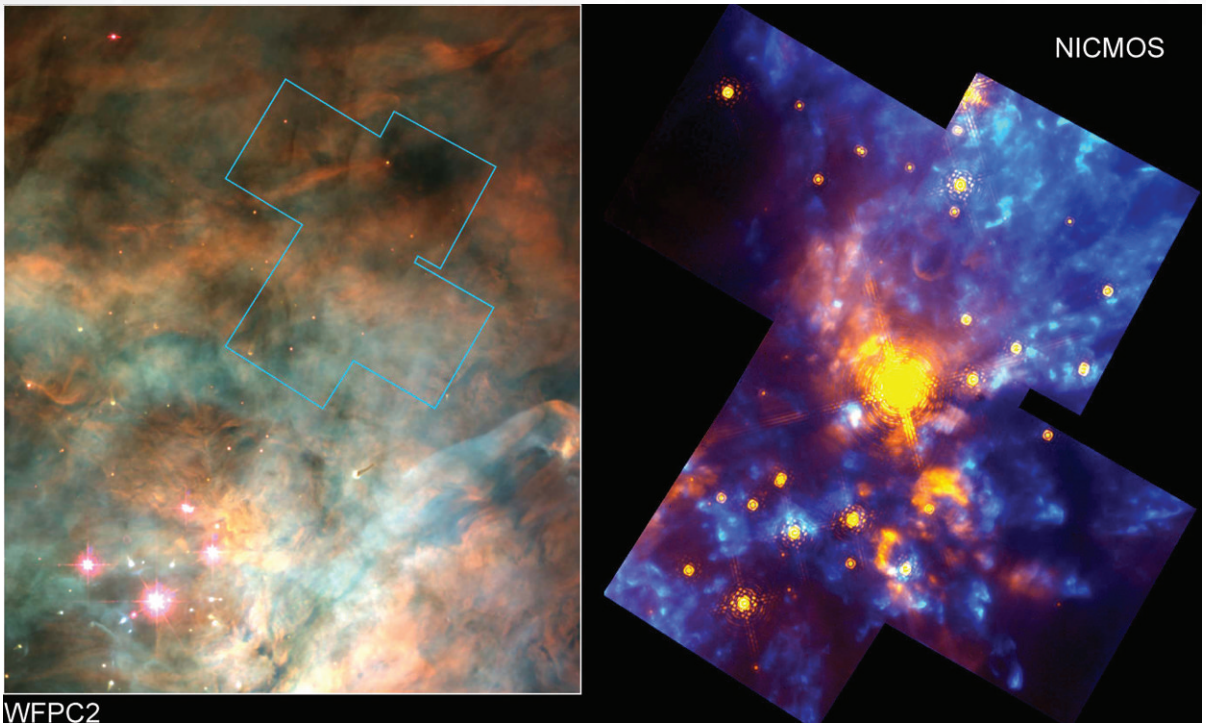


Figure 10.2 (Left) Sharpest image ever taken of the Orion Nebula, where star formation is occurring in a complex tapestry of environments of differing temperature and density some 1,300 to 1,500 light-years from Earth. In the bright central region of the image, called the trapezium because of the arrangement of stars there, ultraviolet light from massive stars is carving out a cavity in the nebula and possibly disrupting star formation there. Image taken by a team led by Massimo Robberto using the Advanced Camera for Surveys on the Hubble Space Telescope. (Right) Hubble near-infrared image of the boxed region reveals newly forming stars hidden by dust in the left-hand panel. NICMOS image by Rodger Thompson, Marcia Rieke, Glenn Schneider, Susan Stolovy (University of Arizona); Edwin Erickson (SETI Institute/Ames Research Center); David Axon (STScI); and NASA. WFPC2 image by C. Robert O'Dell, Shui Kwan Wong (Rice University), and NASA.

Figure 10.4 Hubble Space Telescope image of a jet of material ejected from a disk of gas and dust surrounding a newly formed star. The star is hidden in the lower left portion of the image behind a disk of gas, dust and associated debris. The jet stretches outward trillions of kolometers from the start. This Wide Field and Planetary Camera-2 image courtesy of NASA and the Space Telescope Science Institute.

Figure 10.6 Hubble Space Telescope image of the Eagle Nebula, showing clearing of the gas and dust by ultraviolet light from newly formed stars, which are illuminating the scene. The image was taken by Jeff Hester and Paul Scowen of Arizona State University using the Wide Field and Planetary Camera 2 at visible wavelengths. Courtesy of NASA and the Space Telescope Science Institute.

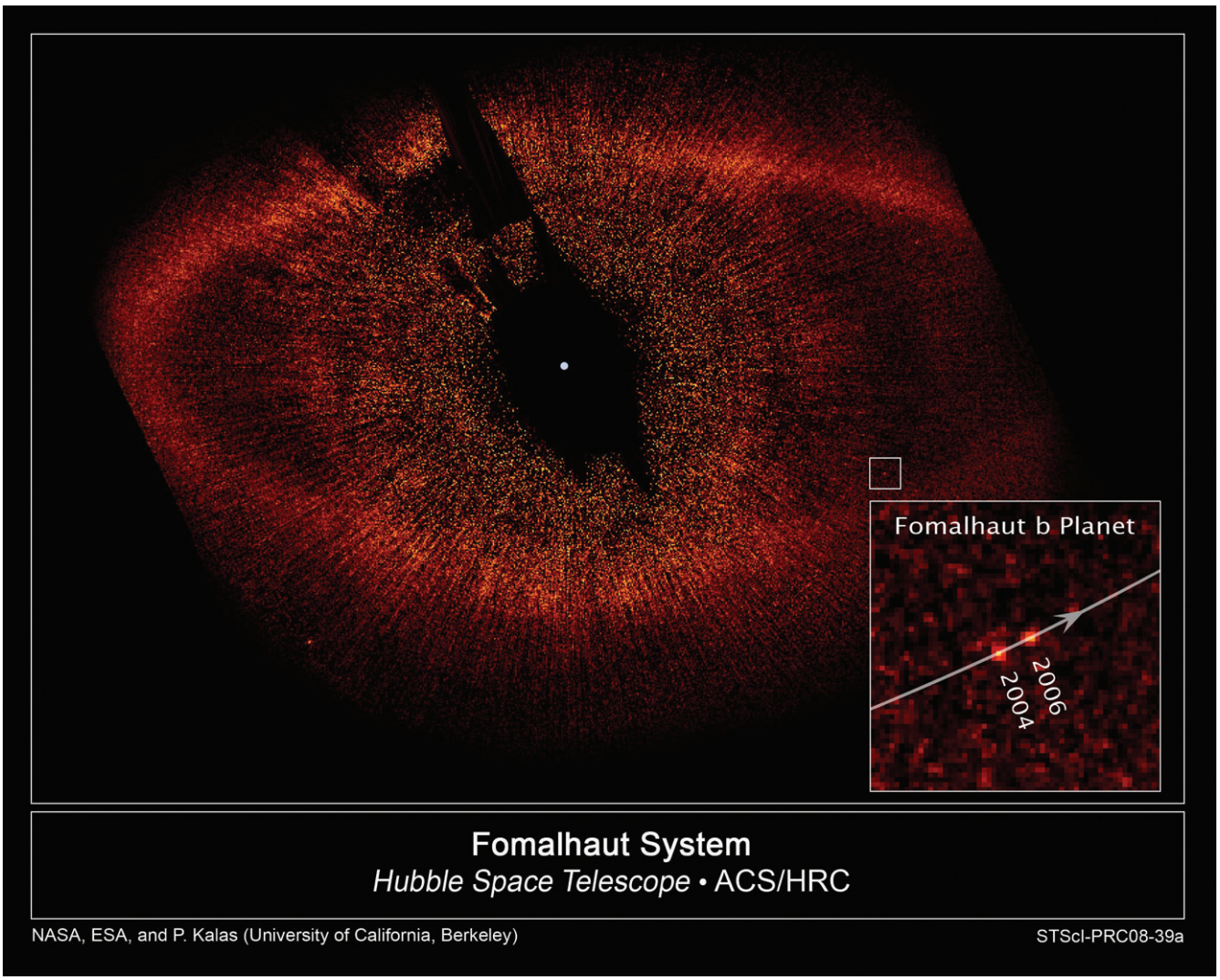


Figure 10.12 A disk of dust is seen around the star Fomalhaut at optical wavelengths, using the coronagraph onboard the Hubble Space Telescope to dim the light of the parent star. The inset is a composite image showing the location of a planet orbiting the star, seen in 2004 and 2006 relative to Fomalhaut. By looking at two succeeding years the motion of the planet can be detected, and indeed it seems to be moving in an orbit nested within the dust belt. From Kalas *et al*. (2008).

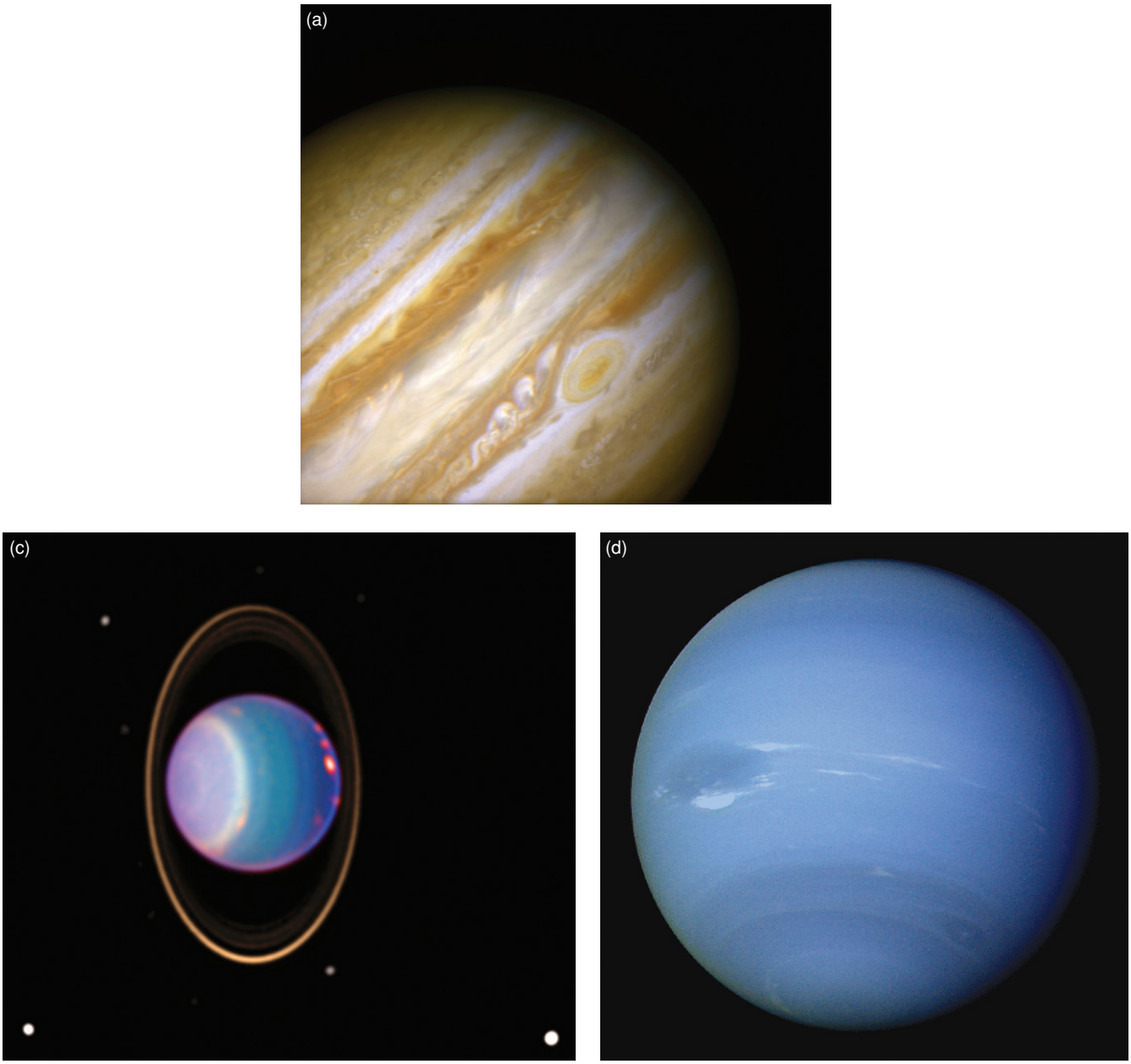


Figure 11.2 The giant planets of our solar system: (a) Jupiter from Hubble Space Telescope; (b) Saturn from Hubble, with contrast exaggerated to show atmospheric patterns; (c) Uranus from Voyager 2, also with enhanced contrast to show very faint banding; (d) Neptune from Voyager 2. Photos (a) and (b) courtesy of NASA and the Space Telescope Science Institute; (c) and (d) courtesy NASA and the Jet Propulsion Laboratory.

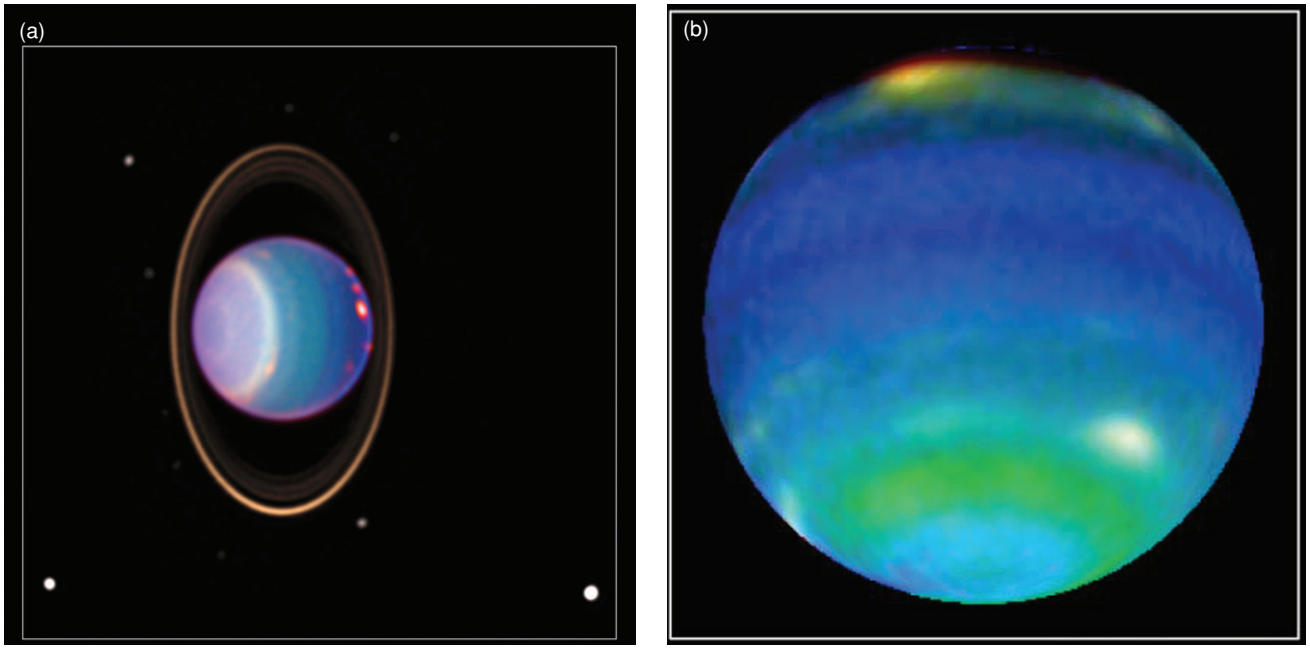


Figure 11.4 Uranus and Neptune from the Hubble Space Telescope, showing continued weather on Neptune (a) and a surge of storms on Uranus (b). Uranus image by Kenneth Seidelman (U.S. Naval Observatory); Neptune images by David Crisp at NASA's Jet Propulsion Laboratory and Heidi Hammel at Massachusetts Institute of Technology Courtesy of NASA and Space Telescope Science Institute.

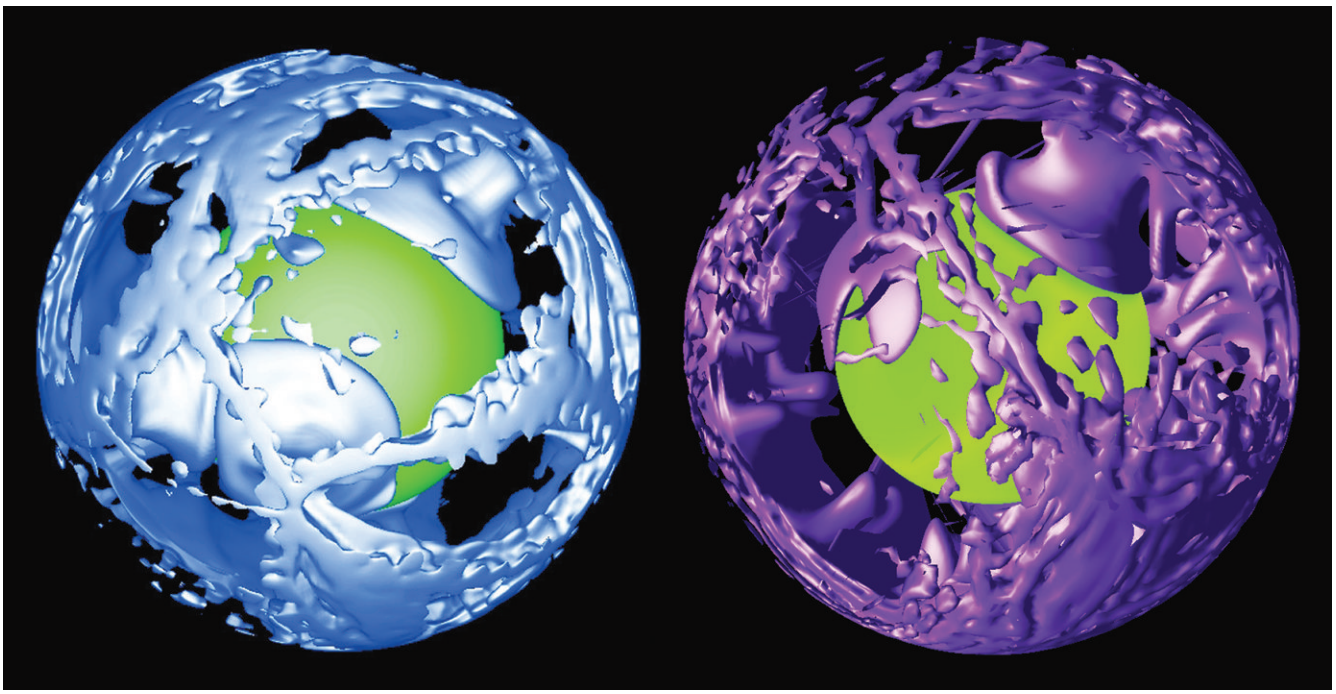


Figure 11.9 Computer model of convection in Earth (Tackley, 1995; Tackley *et al*., 1994). The model is three-dimensional and includes the presence of the phase transition at the upper–lower mantle interface. The left panel shows hot upwelling currents; the right panel shows cold downwelling currents. The inner sphere, which can be partly seen through the mantle currents, indicates the boundary with the iron core, which convects separately. Figures courtesy of Paul Tackley, University of California at Los Angeles.

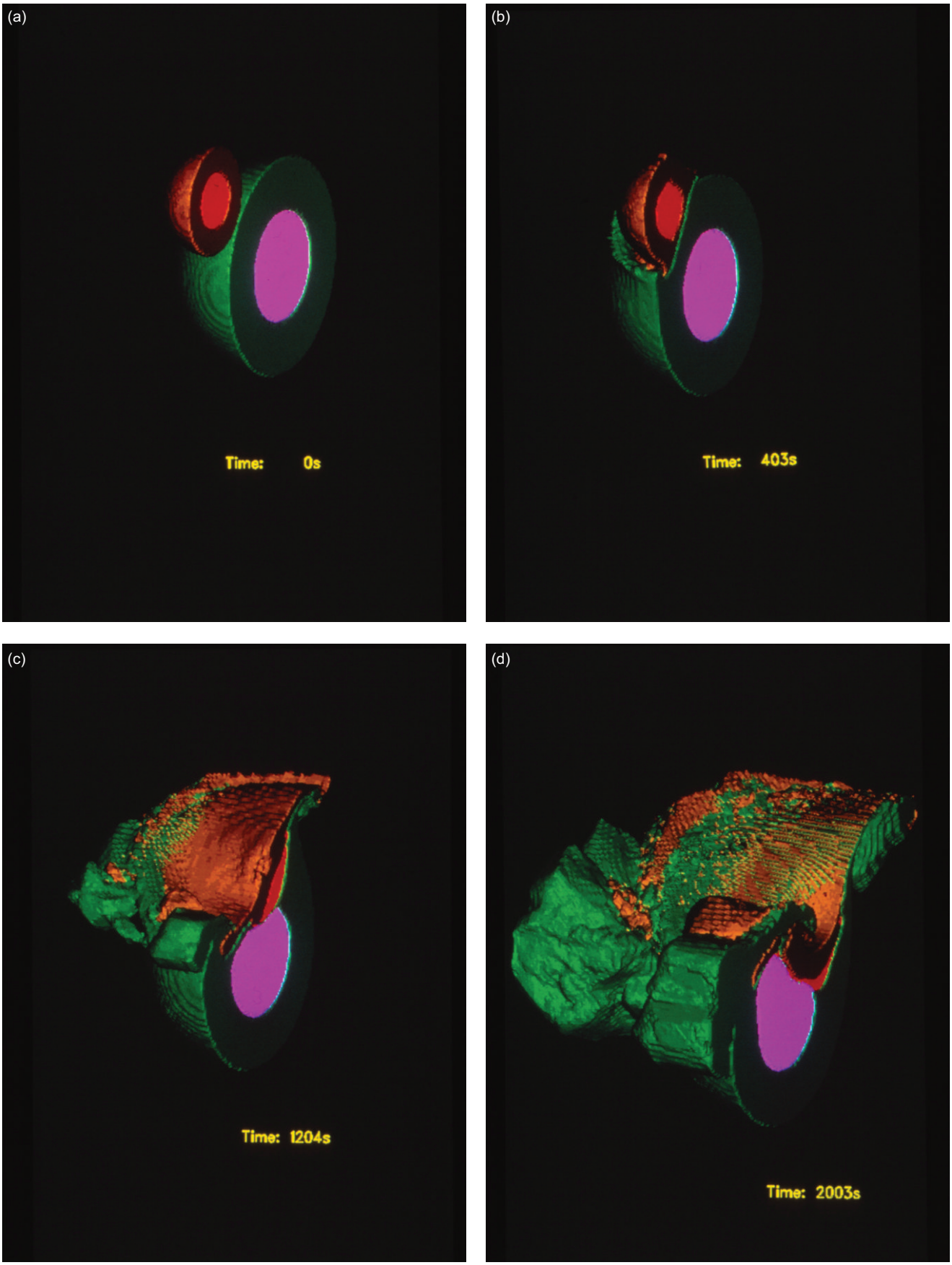


Figure 11.10 Computer calculations by M. E. Kipp (Sandia National Laboratories) and H. J. Melosh (The University of Arizona), showing early stages in the formation of the Moon as a Mars-sized planet strikes Earth. Both Earth and the impacting planet are shown sliced in half so as to reveal what is happening in the interiors. The iron-rich core can be seen as an inner circle in each planet prior to impact. Compared to the mantle of Earth, the core is hardly disrupted. Elapsed time is shown on each panel. Images courtesy of H. J. Melosh.

Figure 13.1 Stellar nucleosynthesis is a primary produces of entropy in the modern universe (a), and provides a source of so-called “free energy," which keeps planetary surfaces away from equilibrium. As Earth is warmed by the Sun and radiates energy in the form of heat, it too generates entropy, but life that is hosted on its surface, considered without regard to its environment, would seem to have decreasing entropy with time (b). When considered as a system coupled to its immediate environment ("env"), however, one sees that the net change in entropy of life plus its immediate environment is positive. (c). Even life existing in the crust of the Earth, heated by radioactive decay of the elements, uses energy that at its source comes from the fusion of elements in stars, since it is there the elements are made. Figures from C. Lineweaver.

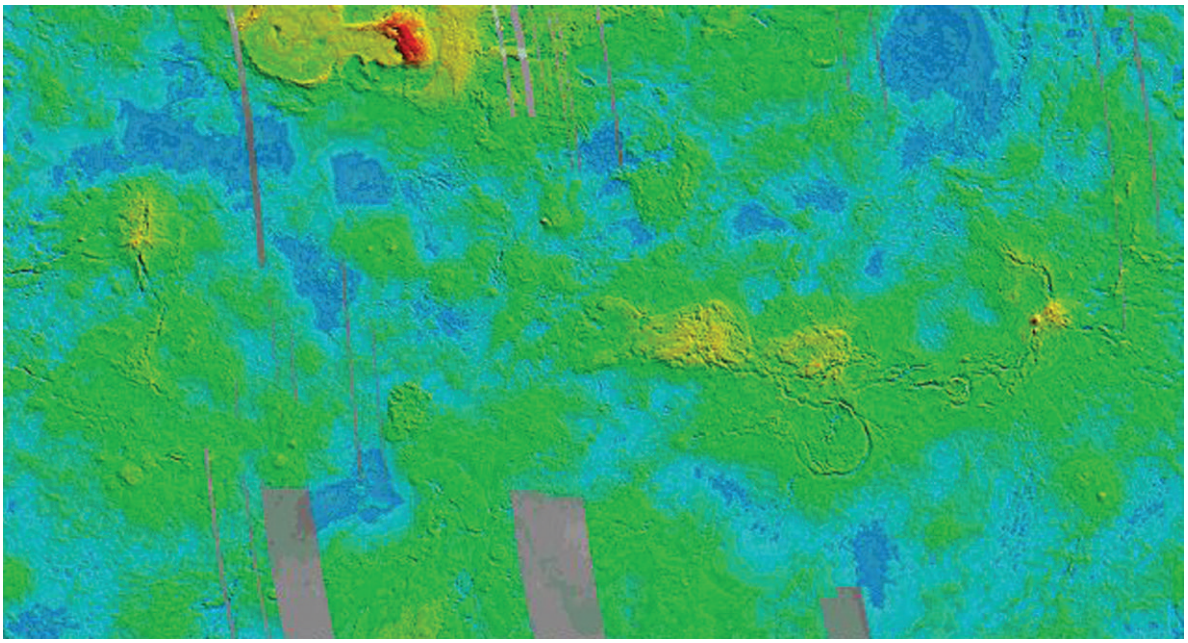
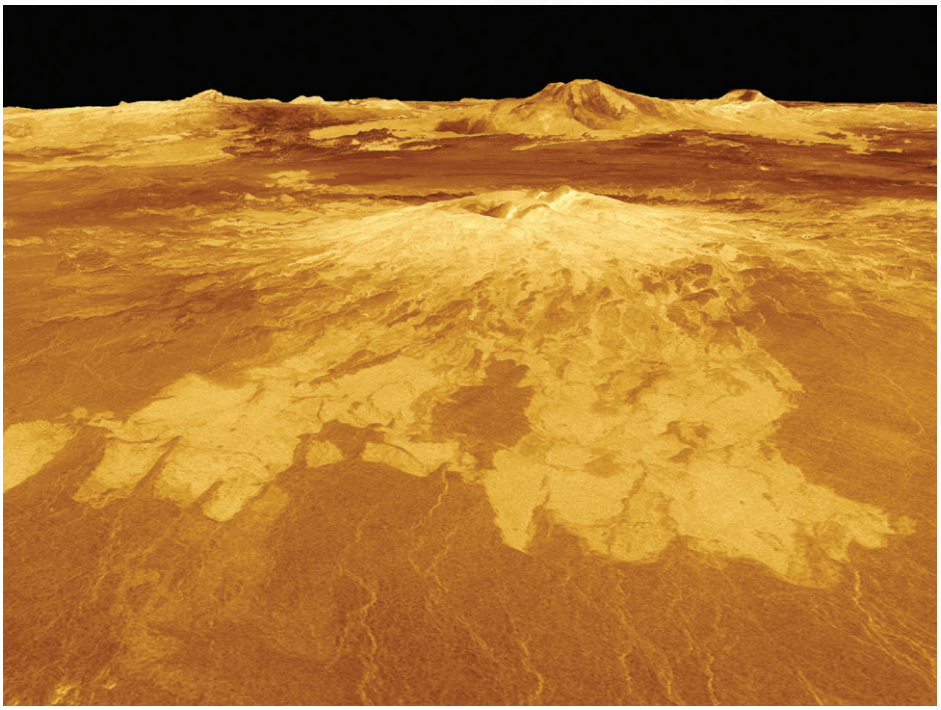
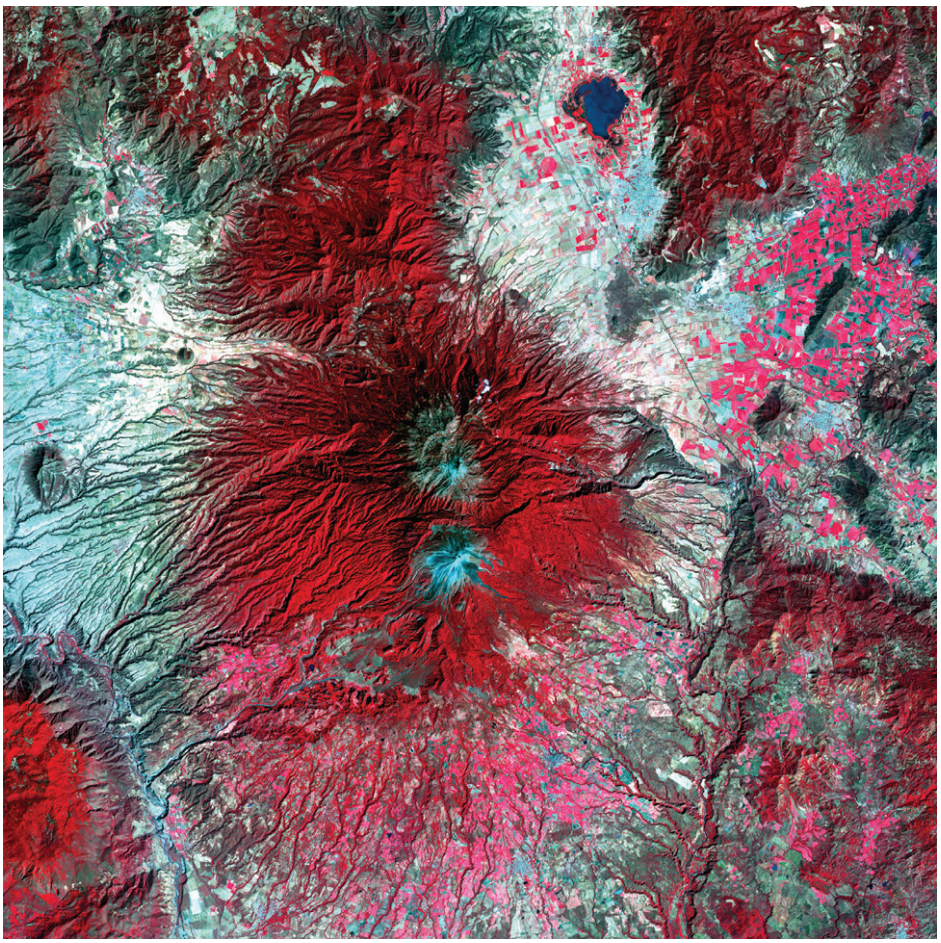


Figure 15.3 Global topography of Venus. Red areas are highest, blue lowest. Courtesy of NASA/Jet Propulsion Laboratory.



(a)



(b)

Figure 15.4 (a) Sapas Mons, a 600 km diameter, 1.5 km high volcano on Venus, shows no evidence of water erosion; the bright linear features have the form and appearance of lava channels. This *Magellan* radar view exaggerates the vertical extent by a factor of 10. Image courtesy of NASA/Jet Propulsion Laboratory. (b) Snow-capped Colima Volcano in Mexico. The southern caldera has been active historically. Calderas and flanks show an intricate network of water-carved channels. The image was made by the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) aboard NASA's Terra satellite. Courtesy of USGS.

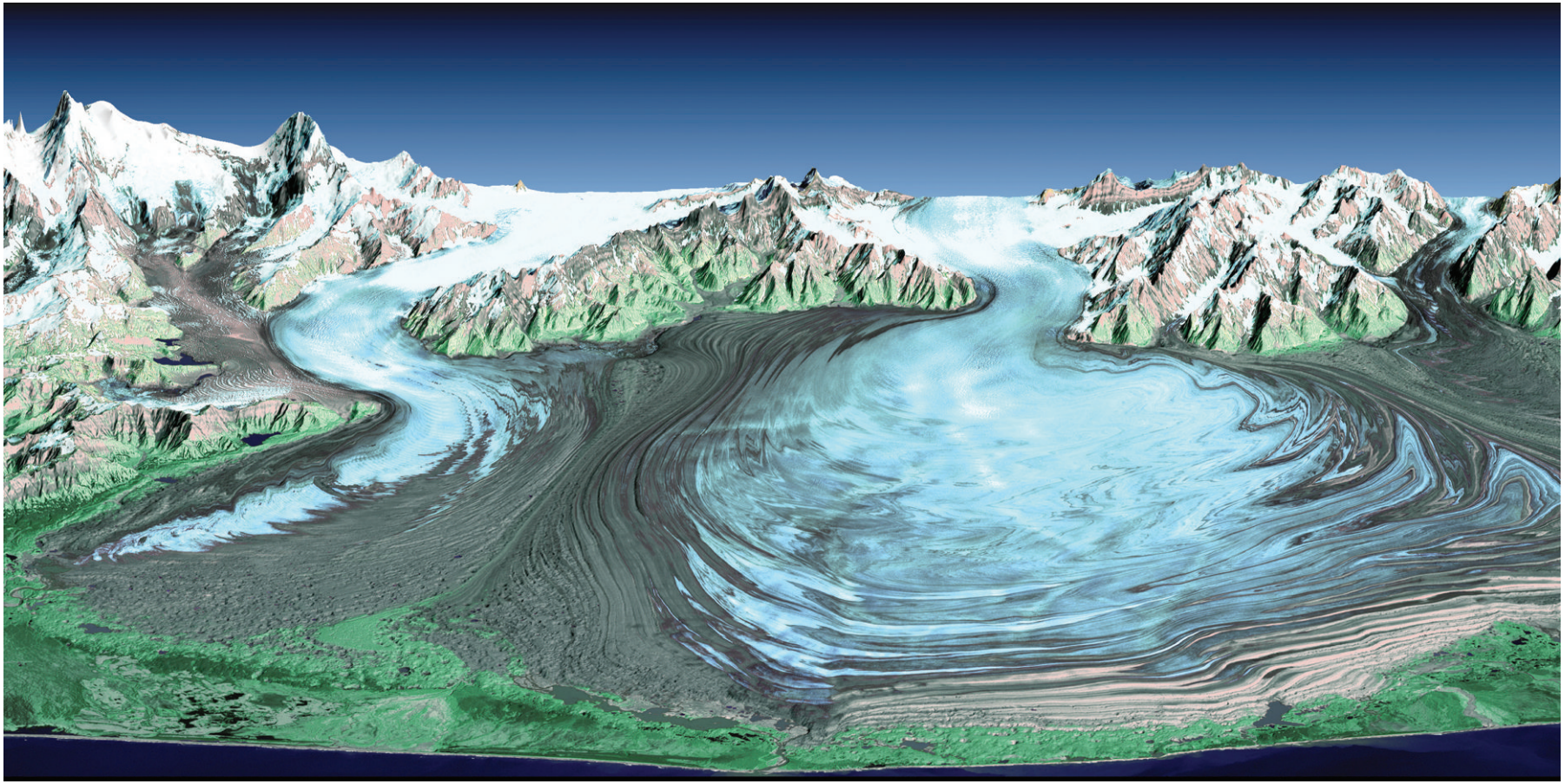


Figure 15.6 Examples of unusual Martian features interpreted to be glacial in origin. (a) Scour marks in Kansei Vallis, appear to be due to glacial erosion rather than by water erosion. The youthful nature of this area suggests that the glacial activity may have been recent. *Mars Express* image from ESA/DLR/FU Berlin (G. Neukum). (b) Piedmont lobe, about 3 km across, seen in Northern Arabia Terra. Such lobes are glaciers flowing out of a confined valley into a broad plain. Image from the Themis instrument aboard *Mars Odyssey*, from NASA/ASU (P. Christensen). (c) A terrestrial equivalent, the Malaspina Glacier in Alaska, is actually the merger of several glaciers. Landsat thematic image, courtesy SRTM Team NASA/JPL/NIMA.

Figure 19.10 Layered deposits of dust and ice at the south pole of Mars.

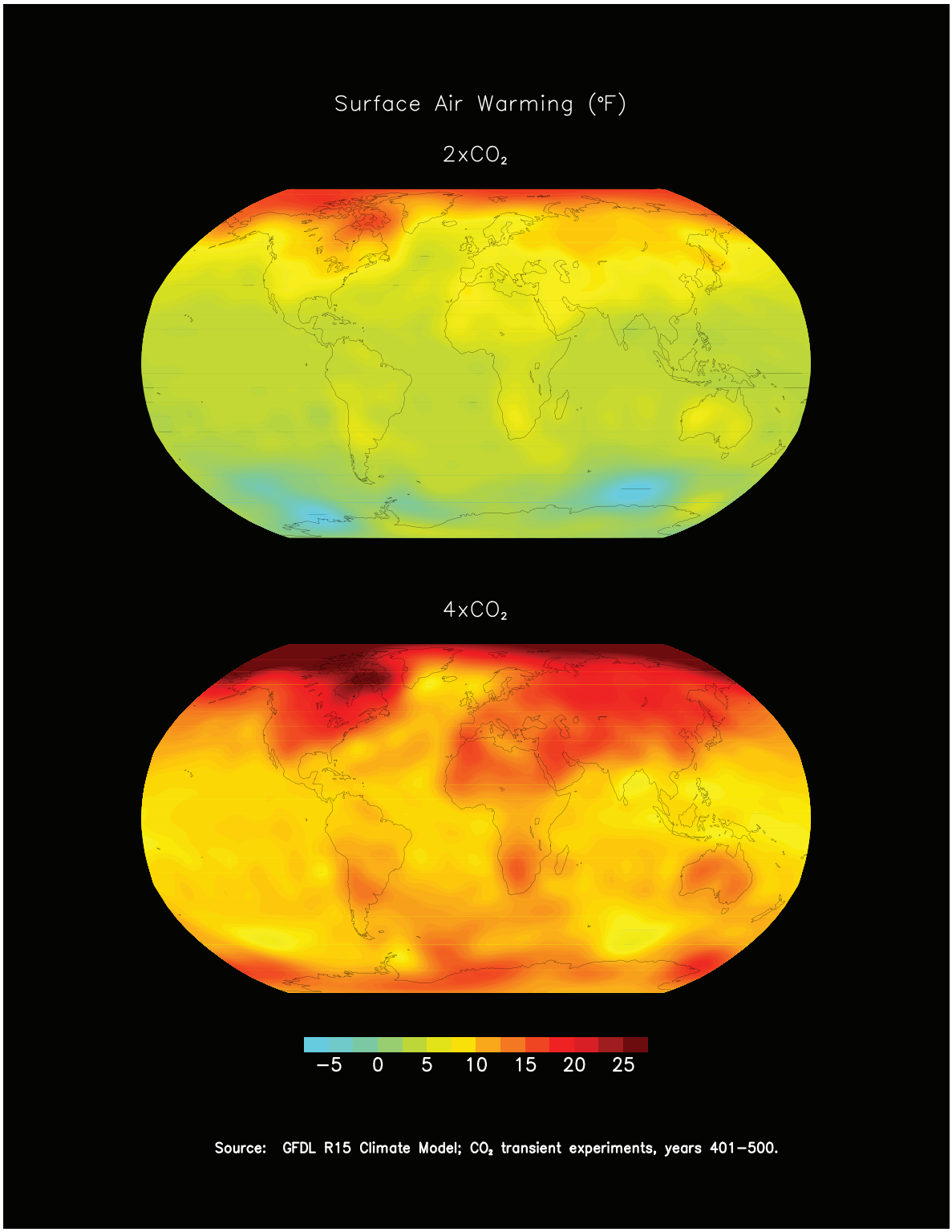


Figure 22.5 Predicted rise in surface air temperature for CO_2 doubled and quadrupled, based on a climate model developed at the Geophysical Fluid Dynamics Laboratory of Princeton University. Temperature is shown in degrees Fahrenheit. Figure created by Thomas Knutson, provided courtesy of the Geophysical Fluid Dynamics Laboratory, National Oceanic and Atmospheric Administration.

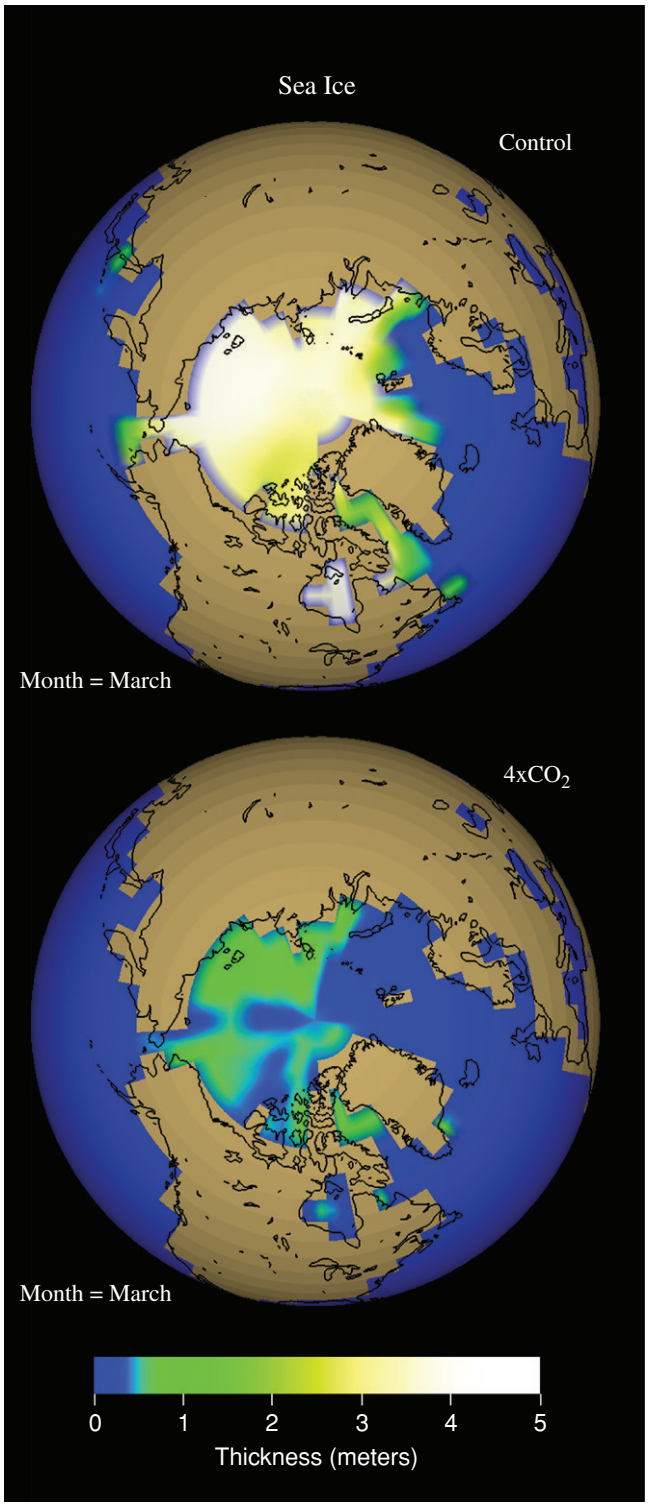


Figure 22.6 Changes in sea ice thickness for quadrupled CO_2 (bottom) relative to no change in carbon dioxide (top panel). Figure created by Hans Vahlenkamp and Thomas Knutson, provided courtesy of Geophysical Fluid Dynamics Laboratory, National Oceanic and Atmospheric Administration.